CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

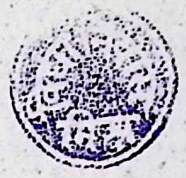
Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

2-1



Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TYPE-TOKEN MATHEMATICS

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

JANUA LINGUARUM

STUDIA MEMORIAE
NICOLAI VAN WIJK DEDICATA

edenda curat

CORNELIS H. VAN SCHOONEVELD

LEIDEN

SERIES MAIOR
IV

1960

MOUTON & CO · 'S-GRAVENHAGE

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TYPE-TOKEN MATHEMATICS

A TEXTBOOK OF MATHEMATICAL LINGUISTICS

by

GUSTAV HERDAN

LECTURER IN STATISTICS
UNIVERSITY OF BRISTOL

डा० विद्या निवास मिश्र
प्राध्यापक, भाषा विज्ञान
निजी पुस्तकालय
पुस्तक संख्या.....

1960

MOUTON & CO · 'S-GRAVENHAGE

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

© Copyright Mouton & Co., Publishers, The Hague, The Netherlands.

No part of this book may be translated or reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publishers.

वा० वि० निवास मिश्र
प्राध्यापक संस्कृत विभाग
निजी पुस्तकालय
पुस्तक संख्या.....

Vidya Niwas Misra
Sanskrit Department
University of Gorakhpur,
GORAKHPUR.

Printed in The Netherlands by Mouton & Co., Printers, The Hague.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

PREFACE

It is said that a writer who understands his subject should be able not only to write a compendium of it, but also to present its essence in a few sentences. I do not know whether this book could claim the name of a compendium of mathematical linguistics, but I am confident that I could give the gist of the matter in the Preface.

There are two plain facts about language which have given rise to the subject of this book: one is that the masses of linguistic forms needed for the manifestation of thought in speech and writing – quite apart from their 'content' belonging to the realm of the mind – are a part of the physical universe, and as such are subject to the laws which govern mass assemblies of any kind. The linguist who wishes for particular reasons to make use of that aspect of language must acquire some knowledge of the laws governing mass events in general and, in particular, the mass events of language. This is how the need for statistical linguistics arises.

An equally undeniable fact about language is that linguistic expression, be it in speech or writing, is a linear sequence of linguistic elements – "*Tout le mécanisme de la langue en dépend*" (de Saussure) – and, insofar, subject to the laws which govern linear arrangements in general. The linguist who wishes for certain reasons to make use of that property of language, namely the linearity of linguistic sequences, must again acquaint himself with the laws governing such arrangements. This is how the need for language geometry arises.

Language in the mass and *language in the line* are special aspects of language, and far be it from me to maintain that they must enter every type of language work. But they do so occasionally and certain branches of linguistic research have been known to profit greatly by taking these aspects into account.

I wish to thank the following professional bodies, editors and authors for their permission to reproduce the diagrams and tables mentioned below: The Royal Society of London (Appendix F); B. Bloch, as editor of *Language*, and C. D. Chrétien, F. Harary and H. H. Paper (Figs. 9, 10); J. Aitchison and J. A. C. Brown, and the Cambridge University Press (Fig. 4 and App. A1); F. Chayes and J. Wiley and Sons, Inc. New York (Figs. 14, 15); W. Feller and J. Wiley and Sons, Inc. New York (Fig. 16, 17); J. S. Huxley and Methuen & Co., Ltd. London (Fig. 1); A. Previtera

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(Appendix E); H. Spang-Hannsen and Mouton & Co., The Hague (Fig. 12); and D. B. Fry, E. S. Pearson (Cambridge University Press), and E. Pulgram (Mouton & Co., The Hague) for permission to reproduce parts of previous publications of mine. I especially wish to acknowledge my indebtedness to R. Morgenthaler, whose book on statistics of New Testament Greek I had the opportunity of seeing in the proof stage, and thus in good time to use some of its numerical material.

For patient help with the laborious work of proof reading and checking the numerical tables, I am greatly indebted to my wife, Innes Herdan; for checking the mathematical formulae to R. Henstock of the Department of Mathematics, University of Bristol.

Bristol, October 1959 G. HERDAN

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

CONTENTS

INTRODUCTION 13

I
STYLO-STATISTICS

I. THE SOURCES OF LANGUAGE STATISTICS AND BASIC CONCEPTS 21
1.1 The element of statistical frequency in language 21
1.2 The sources of language statistics 21
1.3 The statistical frequency dictionary 22
1.4 The logarithmic TYPE/TOKEN ratio 26
1.4.1 Previous observations on the logarithmic Type/Token ratio 26
1.4.2 The law of relative growth of literary products 27
1.4.3 The constant differential growth-rate as a style characteristic 29
1.4.4 Discussion of method 32
1.5 The linguistic frequency distribution 33
1.6 A critical examination of the Zipf law 35
1.7 In defence of the "statistical hypothesis" for the use of words . 38

II. THE DISTRIBUTION LAW OF LINGUISTIC VARIABLES 42
2.1 Lognormality of the word-count distribution 42
2.2 The measure of concentration for word counts 47
2.3 Estimating the word count distribution on the basis of the lognormal hypothesis 50
2.4 Sentence length as a style characteristic 54
2.4.1 Sentence length as a statistical (lognormal) variable . . . 54
2.4.2 Sentence length and authorship 55
2.4.3 Language-conditioned elements of sentence length 57

III. THE STATISTICAL VERSION OF DISCOURSE ANALYSIS—THE DISTRIBUTION OF VOCABULARY OVER SEVERAL SAMPLES 59

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

3.1 Material and notation 60
3.2 Features of vocabulary structure 61
3.3 Random partitioning of vocabulary 62
3.4 How real a phenomenon, or only apparent, is the Hapax Legomenon? 66
3.5 Richness (extent) of vocabulary 68

IV. The Relativity of Vocabulary Ratios 70
4.1 The dependence of the Romance vocabulary ratio upon text length 70
4.2 Application to dating of literary texts 73
4.3 Relative richness of vocabulary 75
4.4 Theoretical derivation of the percentage Romance vocabulary in works by Chaucer 75
4.5 Discussion of results 79
4.6 The relativity of vocabulary ratios 81
4.7 Cause and effect in language 82

V. The Characteristic Constant of the Type-Token Relation . . . 83
5.1 Different interpretations of the characteristic constant K . . . 83
5.2 The geometrical interpretation of the repeat rate 83
5.2.1 The restrictive condition for the geometrical interpretation of the repeat rate 86
5.3 The relation between v_m and the repeat rate 86
5.4 The stability of v_m in translations 88
5.5 v_m as a characteristic for distributions of linguistic forms other than words 90
5.6 The effect upon K of truncating the frequency distribution . . . 90
5.7 The mathematical relation between Greenberg's index of linguistic diversity and Yule's characteristic 92
5.7.1 Greenberg's constant 92
5.7.2 Yule's constant 93
5.7.3 The repeat rate 94
5.7.4 Illustration 94

VI. Vocabulary Correlation 96
6.1 Conventional measures of vocabulary relationship 96
6.2 Shortcomings of conventional methods for establishing linguistic relationship 97
6.2.1 Illustration—the Melanesian language of South Eastern Papua 98
6.2.2 Establishing linguistic relationship by χ^2 99
6.2.3 Special vs. general components of linguistic relationships 100

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

6.3 Vocabulary connectivity and its use for establishing linguistic correlation 104
6.4 Vocabulary connectivity vs. random partitioning 109

II
STATISTICAL LINGUISTICS

VII. The Different Types of Language Statistics 115
7.1 Random sampling in linguistics 115
7.2 Vocabulary and occurrence statistics 117
7.3 The distribution of phonemes with regard to word structure—observed vs. possible frequency of phonemes 119
7.4 Phonological statistics of vocabulary—functional burdening and utilisation of phonemes 121
7.5 A general calculus of phoneme distribution in the dictionary . . 123
7.6 The frequency distribution of phonological oppositions 125

VIII. The Relation Between the Functional Burdening of Phonemes and their Frequency of Occurrence 127
8.1 The phoneme distribution of English according to the dictionary 127
8.2 The utilisation of English phonemes in speech 128
8.3 The derivation of the occurrence distribution from that in the dictionary 130
8.4 Conclusions 131

IX. The Relation Between the Vocabulary Distribution and the Occurrence Distribution of Word Length 133
9.1 Lognormality of the word length distribution 133
9.2 Distribution of occurrences as moment distribution of the vocabulary distribution 135
9.3 The relation between occurrence and lexicon distribution in the light of information theory 138

X. Selective Variation in the Vowel-Consonant Sequence of English and Russian 141
10.1 Material and method 142
10.2 Connected and disconnected letter occurrence 142
10.3 The Markoff chain mechanism 146
10.4 Comparison of Russian and English vowel-consonant sequences 150
10.5 Markoff chain and mechanical translation 152

XI. The Repetitiveness of Linguistic Forms 154
11.1 The distribution of gaps between the repetitions of linguistic forms 154

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

11.2 The repeat rate 156
11.3 Statistical analysis of language contact—the repeat rate of initials 158
11.3.1 Language mixture. 158
11.3.2 Loanwords. 164

III
INFORMATION THEORY

XII. The Coding of Language in the Brain 171
12.1 The linguistic interpretation of information, entropy, redundancy 171
12.2 The units of information 173
12.3 Transmission of information using a binary code 174
12.4 Entropy as a measure of information. 176
12.5 A different derivation of the entropy as a measure of information. 177
12.6 The storage of information in the brain 179

XIII. The Mathematical Condition for Optimal Redundancy 182
13.1 Efficiency of artificial codes—Pitman's Shorthand as a case of optimal coding 182
13.2 Generalisation of the condition for optimal coding 184
13.3 The uncertainty relation of information theory 186
13.4 H/H_{max} as criterion of the quality of coding 186
13.5 Binary coding of words as a case of optimal coding 188
13.6 Optimal coding and Fibonacci numbers 189

XIV A. An Inequality Relation Between Yule's Characteristic K and Shannon's Entropy H 192
14.1 Entropy and linguistics. 192
14.2 A transformation of the entropy—the inequality relation between K and H 192
14.3 Application (uni- and bi-variate) 195
XIV B. Markoff-chain mechanism in language translation 196
14.4 Theory. 196
14.5 Application. 199

XV. The Skew Distributions of Statistical Linguistics. 201
15.1 Reciprocity relations of linguistic forms 201
15.2 Artificial reciprocity relations 202
15.2.1 The distribution of frequency according to symbol length 202
15.2.2 The distribution of frequency according to rank 203

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

15.3 Natural reciprocity relations 206
15.3.1 Relations between some important skew distributions of linguistics 207
15.3.2 The inadequacy of Simon's assumptions 208

XVI. The Segmentation of Meaning 210
16.1 Structure in language 210
16.2 Classification as a stochastic process 211
16.3 The probability mechanism for the creation of new concepts—I. Time assumed infinite 213
16.3.1 The bi-logarithmic straight line law 217
16.4 The probability mechanism of creating new concepts—II. Time assumed finite 217
16.5 Illustration 218
16.6 How concept content changes with concept size 220
16.7 The Rosin-Rammler law applied to the reduction of content with increase of size of concepts 221
16.8 Aspects of quantitative language structure 225

IV
LINGUISTIC DUALITY

XVII. The Type-Token Duality 229
17.1 The principle of geometrical duality in language—interchangeability of type and token in linguistic statements 230
17.2 Duality of linguistic theorems 232
17.3 Linguistic duality and uni-laterality of the speech function in the brain 233
17.3.1 Bi-lateral and uni-lateral brain functions 234
17.3.2 The doubling mechanism of language 235
17.3.3 The doubling mechanism of language compensates for the uni-laterality of the speech function 238
17.4 Linguistic duality as a methodological tool of research 239
17.5 The Distributionalist—Prague School controversy in the light of linguistic duality 240

XVIII A. The Type-Text Duality 243
18.1 The Type-Token-Text relation as a case of geometrical duality in three dimensions 243
18.2 Partitioning of vocabulary—unity of style vs. unity of content 245
18.3 Distribution of vocabulary in the New Testament 245
18.4 Discussion of results—the structure of meaning 249
18.5 Random partitioning of vocabulary as a standard of comparison 252

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XVIII B. The Structure of Meaning—The Prosodic Properties of Prose Diction 255
18.6 The dynamic view of vocabulary partitioning—the structure of meaning 255
18.7 The type-content relation as a case of geometrical duality in four dimensions 257
18.8 Frequency of occurrence as periodicity in the use of words . . 258

XIX. Mathematical Philology 262
19.1 Language about language 262
XIX A. The Craft of Satire 263
19.2 The decay of language through the abuse of linguistic duality—a linguistic analysis of Nazi German 263
19.3 The craft of satire 264
19.4 The power of linguistic duality 266
XIX B. The Craft of Fiction 270
19.5 Drama vs. fiction 270
19.6 The craft of fiction—duality in five dimensions 271
19.7 Philological duality 272
19.8 Plurality of codes as a device of the craft of fiction 273
19.9 Summary 274
19.10 Illustrative material 275
19.10.1 Tolstoy, "War and Peace" and "The Kreutzer Sonata" 275
19.10.2 Theodor Storm, "Der Schimmelreiter" 276
19.10.3 Goethe, "Iphigenie" and "West-Östlicher Divan" . . 277
19.11 The linguistic duality of poetic diction 278

V
THE CALCULUS OF LINGUISTIC OBSERVATIONS

XX. Mathematical Characteristics of Language Structure 281
20.1 An exordium on linguistic sampling 282
20.1.1 Equivalence of the proportions obtained by sampling counts and the proportions in the whole of a literary text 282
20.1.2 Reproducibility of estimates of proportions of linguistic forms 286
20.1.3 Illustration from stylostatistics 287
20.1.4 Sampling efficiency of linguistic material 291
20.2 The universes and population probabilities of literary statistics 292
20.2.1 The universes of literary statistics—The universe of content 292
20.2.2 The linguistic population probabilities 293

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

20.3 The ergodicity of word masses 295
20.4 A characteristic quantity of linguistic statistics—the repeat rate 297
20.5 The sum of least squares of language statistics 298
20.5.1 Conditioned repeat rates 300
20.6 The binary principle of coding in the light of geometrical probability 301
20.7 The statistics of the structure of meaning 303
20.8 Logarithmic probabilities and logarithmic variables 307
20.9 Projective geometry as the mathematics of non-measurable properties of language 308
20.10 Specificity of parameters and methods 312

XXI. The "Eigenstatistics" of Language 314
21.2 Epistemological nature of the principle of linguistic duality . . 314
21.2 Statistics from "without" and "within" 315
21.3 The phonological opposition 315
21.4 The "Eigenfunction" of linguistic control—Language as having compensatory devices for the one-sidedness of the speech function in the brain 316
21.5 The structure of meaning—the incomplete balanced block design of vocabulary structure 318
21.5.1 The system of controls in experimental design according to the analysis of variance 318
21.5.2 The distribution of vocabulary in literary samples in the light of the principles of experimental design 319
21.5.3 The structure of meaning as a statistical invariant . . . 320

Epilogue 323

Mathematical Appendices 333

Glossary of Statistical Terms 413

Glossary of Linguistic Terms 425

Bibliography 435

Index 443

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

NOTE

Numeration of Sections, Tables, Figures, and Equations

Each section in this book is distinguished by a number consisting of the chapter number prefixed to the number of the section in the chapter, e.g. 7.5 means the 5th section of the 7th chapter; sub-sections are distinguished by a second digit after the period. Equations are numbered separately in each Part; tables and figures are numbered consecutively right through the book.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

INTRODUCTION

This volume is a continuation of the author's *Language as Choice and Chance* and develops the subject by presenting new methods applied to new material, arranged according to the same sub-divisions as in the first volume. All efforts have been made to make the contents accessible also to readers who have not read the older book; terms used there are defined and important formulas repeated.

It is, however, also a new book, insofar as it stresses the peculiarity of statistics of language structure as against just literary statistics. To put the difference between the two types of statistics briefly: the latter uses the methods and parameters of general number statistics as applied e.g. in economics and demography, the former has its own methods and characteristic parameters particularly useful for describing and evaluating language structure. The idea of statistical linguistics as using concepts and methods of its own, which was adumbrated in the first volume has now taken shape (Part I, II). Analogously, in Part III linguistic information theory is shown to be something different from a straightforward application of certain concepts of statistical physics (such as the Entropy), which leads, among other things, to establishing the mathematical conditions for optimal coding (Chpt. XIII).

For the presentation of the statistics of language structure there seemed to be two courses open. One was to use as a starting point the theory of probability or of mathematical statistics, and to show how through certain modifications in the fundamental assumptions the theory could be adapted to the work on linguistic data, and how, in this way, a new type of statistics arises.

The second way is to take as the starting point certain problems of structural linguistics and to show language statistics as comprising the mathematical procedures by which these problems could be successfully tackled.

The first course is not one to recommend itself in a book of this kind, since it presupposes on the part of the reader a knowledge of mathematical statistics which could not be generally assumed among linguists. Therefore, the second course was adopted.

In accordance with this plan, mathematical proofs of statements made in the text and requiring such proof are collected in an Appendix.

This choice of method of presentation does not, however, mean that the comparison

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

with the other established branches of statistics was abandoned. To have done so would, indeed, be a serious omission, since the reference to those branches is not only very often revealing and helpful to the understanding, but it enables one to stress the specificity of language statistics. Apart from occasional references throughout the book, the last two Chapters are devoted to a more systematic exposition of the principles of "The Calculus of Linguistic Observations" (Part V), where the peculiar fabric of this branch of statistics becomes apparent, especially its blending of three different types of statistics: *number statistics* (as stylostatistics and linguistical statistics), *statistical physics* (in information theory), and *quantum statistics* in the form of Bose-Einstein statistics, here introduced for the first time for the study of word content and structure of meaning. This type of statistics is also known as the "New Statistics" to distinguish it from the other types representing the "Classical Statistics".

This leaves Part IV to account for. Information theory as the quantitative expression of de Saussure's attempt to conceive of language as a code of signs has, no doubt, contributed to our realising the difference between the manifold of objects in space and time, or of the corresponding concepts, and the universe of linguistic expression, but it has not done so completely. There is one aspect of language which information theory does not reach. Conceiving of language as a code for the expression of thought implies that thought is primary and language secondary, a means to an end, namely the expression of thought for the purpose of communication. Language as a creative force, and thus as an art, has no place in information theory. Nor has it ever been adequately defined as such, however insistent certain writers may have been in stressing emphatically the difference between language as a means of communication and language as an art. It is this aspect of language which will occupy us in Part IV, where the productive or generating force of language will be shown to be more fundamental for its structure than its communication properties. The following paragraphs give a brief indication of that aspect of language structure.

The word "Mathematics" in the title comprises in addition to statistics also geometry, viz. projective geometry. Geometrical methods—apart from those of coordinate geometry—have tentatively and haphazardly been used in structural linguistics, sometimes without sufficient realisation of what is implied in such procedures. A more fundamental approach is needed to make linguists aware of the role which geometry is to play in structural linguistics, and beware of misinterpretations which may follow in the wake of regarding language structure as a sort of applied geometry.

The justification for the use of geometry in structural linguistics as given here follows these lines:

Into any set of axioms there must enter certain undefined concepts, such as "point" and "line" in geometry. They can be regarded as purely abstract entities whose properties in a deductive system are entirely given by the relations that hold between them as stated by the axioms. For the application, on the other hand, of such a set

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of axioms, it is important that the concepts and axioms of geometry, say, correspond well to physically verifiable statements about "real" tangible objects. Einstein has expressed this as follows: "Euclidian geometry, considered as a mathematical system, is a mere play with empty concepts, straight lines, planes, points etc. being mere fancies. If, however, one adds the sentence that the straight line corresponds to a rigid rod, geometry is transformed into a physical theory."

It has now come to light that if we assign to the "point" of geometry the role of linguistic form (phoneme, morpheme, etc.) which we shall briefly refer to as TYPE, and to the "line" that of linear sequences of such forms (in words, sentences, texts etc.) which we shall refer to as TOKEN, there appears to be, on the whole, agreement between the properties of "types" and "tokens", as the linguistic "points" and "lines", and the formal axioms of geometry. In other words: the statements of geometry remain true in terms of "type" and "token" and geometry is transformed into a linguistic theory.

As a criterion for the correctness of our claim to have made geometry truly applicable to language by letting "type" stand for "point" and "token" for "line", we use the fundamental property according to which the theorems of projective geometry remain true if "point" and "line" are interchanged. This relationship is called DUALITY, and gives projective geometry a character quite distinct from elementary (metric) geometry, in which no such duality exists. This principle asserts for language that if in any valid proposition of language the words TYPE (linguistic form, e.g. phoneme, morpheme) and TOKEN (frequency of occurrence, probability) are interchanged, the resulting proposition is also valid. That statement can be extended so as to comprise the relation between TYPE, TOKEN and TEXT, which leads, among other things, to a new conception of the "structure of content or meaning."

If quantitative methods are useful for describing and analysing linguistic events which are the raw material of literary (philological) achievements, they should also prove useful in describing and analysing the latter. This is actually the case, and here we enter upon a new field of application of mathematics to language: *mathematical philology* (Chpt. XIX). The methods which will be found useful here are not those of the mathematics of measurement (statistics), but those of the mathematics of order and position.

The plan of the book, its system in outline, in terms of Type (*Ty*), Token (*To*), and Text (*Te*) is as follows.

Part I. *Ty* stands here for "word" and *To* for its repeated use. After describing the distribution of *To* according to *Ty*, and its reverse, that of *Ty* according to *To* (Chpt. I, II), the partitioning of *Ty* among a number of *Te* is investigated (Chpt. III). The complicated change of *Ty* with increase in *To* leads to the introduction of the "Characteristic" K and v_m (Chpt. V) as a stable statistic which is independent of the sample size; it can also be described as the average *To*-ratio, in contrast to the *Ty*-ratio which we study in Chpt. IV.

Part II. *Ty* denotes here mostly phonemes, letters, word length and *To*, con-

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

sequently, their occurrence in both language and speech (in the sense of de Saussure's "langue" and "parole") (Chpt. VII). The relation between the distribution of *Ty* in the dictionary and in speech is the subject of Chpts. VIII and IX for phonemes and word length respectively.

Part III. *Ty* may here denote either phonemes or letters or words (qualitative or quantitative characteristics) and *To* is the linear sequence of *Ty* in continuous pieces of a literary *Te*, or of the spoken language. "Information" is shown to be dependent on the solidarity of *Ty* through its categories having definite stable frequencies *To*. Moreover, *Ty* is reduced to its simplest possible system, the binary code, which, among other things, provides a common denominator for the comparison of different systems of *Ty*, (Chpt. XII), and criteria for systems of coding with optimal redundancy (Chpt. XIII).

Part IV. Here a very fundamental property of the *Ty-To* system is discussed: *Duality*. It is shown that what corresponds to the *Ty-To* duality in two dimensions is the *Ty-Te* duality in three with *To* as the connecting link (Chpt. XVII). By way of contrast to the tendency of information theory to substitute the very simplest code for more complicated ones, the use of linguistic duality enables us to go in the other direction, to generalise, and to see the *Ty-To* and *Ty-Te* relation manifested not only on the alphabetic, phonemic, morphemic, vocabulary etc. level, but between pages and texts, books and subjects, individuals (of fiction) and story, and so on. We thus pass from mathematical linguistics to mathematical philology. The integrating property of the duality device counterbalances effectively the analytical tendencies of information theory (Chpt. XIX).

The division of the subject of mathematical linguistics into four parts: stylostatistics, statistical linguistics, information theory, linguistic duality, and thus its system, according to the first volume, appears fully justified in the light of the subsequent development of the subject as described in this book. When dealing with chance variables of the type encountered in language, we must distinguish between three fundamental features:

(1) the structure of the variable (sometimes called the mathematical model) or, as we say in the case of a statistical variable, the structure of the particular universe;

(2) its observable properties, or its manifestation in space and time as "samples" of the universe;

(3) provided these variables are comprised in a larger whole, the relation between the various universes within that whole.

Our subdivision of mathematical linguistics follows such a pattern, and may, therefore, claim to give a complete picture of the matter:

(1)[1] *statistical linguistics* and *stylostatistics* are concerned with what mathematical statisticians call "the sample space" of the various levels of linguistic coding, such as the alphabetic, phonemic, morphemic, vocabulary level, as statistical universes;

[1] Corresponds to (2) above.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(2)[2] the comparison of these universes (in terms of efficiency of coding) with regard to the *same level* of the linguistic coding systems, at different times of development of the same language, or at the same time but in different languages, is made possible by *statistical information theory*, using a binary code as the common denominator for different languages;

(3) the relation between the *different levels* of linguistic coding, and between 'signifiant' and 'signifié', is governed by the principle of *linguistic duality*.

Since a good deal of linguistic controversy, and many misunderstandings of quantitative methods, arise out of the fact that often the problems are not seen in their proper perspective, the reader is advised not to lose sight of the scheme of mathematical linguistics as outlined above when tackling a particular problem.

At the end of the Introduction to the first volume the close relationship between that book and a classic of linguistics: de Saussure's *Cours de Linguistique Générale* was stressed, the work being described as a quantification of de Saussure's fundamental concepts and distinctions. By way of contrast, we end the present Introduction stressing the close relationship between this book and a classic of higher mathematics: G. Boole's *Laws of Thought*. The present work can be described as an extension of Boolean principles from symbolic logic to linguistic expression in all its aspects, the fundamental law in both fields being a law of duality: Boole's algebraic law of duality as the fundamental law of thought, i.e. of *linguistic content*, and our principle of geometric duality in language as the fundamental law of *linguistic expression* (Epilogue).

Glossaries of mathematical and linguistic terms were included to meet the occasional complaints by linguists who were at a loss to understand the mathematical, and by mathematicians, engineers and philosophers who did not understand the linguistic terms. The mathematical glossary is the more comprehensive of the two, owing to the nature of the book, but even so does not aspire to completeness.

For facilitating computations, numerical tables of certain mathematical functions have been added, of which the Random Partitioning Function (App. G), required for establishing vocabulary connectivity, is here tabulated for the first time.

[2] Corresponds to (1) above.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

FIRST PART

STYLO-STATISTICS

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

I

THE SOURCES OF LANGUAGE STATISTICS AND BASIC CONCEPTS

1.1 THE ELEMENT OF STATISTICAL FREQUENCY IN LANGUAGE

Of the various new concepts introduced by quantitative linguistics that of the frequency of occurrence of particular linguistic forms may be said to have been universally accepted by linguists of all schools as an important tool of linguistic research. The reason is that frequency of occurrence has come to be recognised as one of the basic aspects of language.

"Language as the apparatus for the communication of thought within a speech community has by convention three main aspects: Phonemic system, Lexicon and Grammar. By 'members of the speech community' we mean persons using the same phonemic system, lexicon and grammar, disregarding the variations to which each of these tools is subject. However, as quantitative studies have shown, there is another common factor: there exists a far-reaching similarity between members of the speech community, not only as regards the phonemic system, lexicon and grammar they use, but also as regards the *frequency* of use of particular phonemes, lexicon items (words) and grammar forms and structures. In other words: a similarity not only in *what* is used, but also in *how often* it is used."[1]

1.2 THE SOURCES OF LANGUAGE STATISTICS

In accordance with the important role which the statistical frequency of linguistic forms plays in the communication of thought, it would seem desirable to have reliable and complete information about it. In the absence of what may be called "frequency dictionaries", such information can be obtained only by carefully designed and often laborious counts of linguistic forms as they occur in either the depositories of codified language (e.g. dictionaries) or in literary texts.

The advent of electronic calculating machines which are capable of all kinds of lexicographic and word-counting work has changed this. The time is approaching when linguistic material of this type, in sufficient quantity, will be made available to the linguist. His task will then be to analyse that material, and this book (in Parts I-III) provides the necessary tools for such an analysis. There exist certain sources

[1] G. Herdan, *Language as Choice and Chance* (1956), which will be referred to as "Vol. I".

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

from which material can be culled, in more or less the form in which we would require it. Such sources are for our branch of knowledge the various codifications of linguistic forms: dictionaries, grammars, concordances. Of these, grammars are the least helpful; dictionaries are more suitable for our purpose, though they are only means to an end since we have to do the counting ourselves; concordances are the most useful of the three, but even they do not save us the trouble of counting the number of passages in which a particular word or form occurs. Moreover, they are works of rare occurrence, having been compiled for only a small number of writers. (For a comprehensive list of such works, see the *Bibliographie Critique de la Statistique Linguistique*, by P. Guiraud & J. Whatmough, Utrecht, 1953.)

The ideal source would be a statistical frequency dictionary giving for each vocabulary item its frequency of occurrence in the language. In this category fall—to some extent, at least—the frequency lists of words in a language, e.g. E. L. Thorndike and I. Lorge, *The Teachers Word Book of 30,000 words* (for English), and Van der Beke's list (for French). They suffer, however, useful though they are, from the disadvantage of being based upon a sample of the language, for whose appropriateness for the purpose we have to depend upon the author's views as to what is representative of the language as a whole. There are not many lists of this kind which have not been attacked by users on the ground that they do not agree that the sources and method of compilation had been such as to justify the claim of the list being truly representative.

Apart from this, they do not—and cannot, as a rule—show the contribution to a particular word's frequency by the various sources, there being too many. Insofar they are one-sided. A frequency dictionary if it is to comply with this requirement, and avoid all criticism as to representativeness, would have to confine itself to the works by one author—like the concordances—or a more or less homogeneous group of works by different authors.

1.3 THE STATISTICAL FREQUENCY DICTIONARY

Such a work is R. Morgenthaler's statistical dictionary of the Greek New Testament (1958).

Even here, in spite of the restricted sphere of the work, the difficulties were considerable, but the author may be said to have been fully aware of Trubetzkoy's warning: "Man darf sich die Sache nicht zu leicht vorstellen" (Trubetzkoy, 1936). There are the differences in opinion as to which New Testament text should be used for the whole enterprise. The author, having rightly stressed that small differences between texts cannot have any significant effect upon the results of a mass count of ca. 140,000 words, has chosen the 21st edition of the Nestle text. However, he did not depend upon the mere word count according to that text, but decided to check that count by means of concordances. Here again the question arose which concordance to use. The way out of this difficulty was to combine the information from the concordances by Geden and Bruder, and the dictionary by Bauer, and thus to achieve a multiple control. The

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 1

New Testament Text length (No. of words) according to			Vocab-ulary	*Pauline Letters* Text length (No. of words) according to			Vocab-ulary	*Catholic Letters* Text length (No. of words) according to			Vocab-ulary
	Nestle text	Concor-dances by Bruder, Geden and Bauer			Nestle text	Concor-dances by Bruder, Geden and Bauer			Nestle text	Concor-dances by Bruder Geden and Bauer	
Mt	18305	18278	1691	Ro	7105	7094	1068	Ja	1749	1735	560
Mk	11242	11229	1345	1 Co	6811	6807	967	1 Pe	1678	1669	545
Lk	19428	19404	2055	2 Co	4469	4448	792	2 Pe	1098	1103	401
Jh	15416	15420	1011	Gal	2229	2220	526	Jo	2601	2597	302
Act	18382	18374	2038	Eph	2418	2425	529	Ju	457	456	227
Pls	32349	32303	2648	Phi	1629	1624	448				
Hb	4951	4942	1038	Col	1575	1577	431	Total	7583	7560	1271
CB	7583	7560	1271	1 The	1475	1472	366				
Rev	9834	9818	917	2 The	821	824	250				
				1 The	1588	1586	541				
Total	137490	137328	5436	2 Ti	1236	1235	458				
				Ti	658	663	303				
				Phe	335	328	141				
				Total	32349	32303	2648				

Observe the excellent agreement between the text length of the whole New Testament and its parts in terms of number of words obtained directly from a count of the Nestle text, and indirectly from the Concordances.

result is the best possible vocabulary statistics attainable under the circumstances; absolute correctness being neither possible nor necessary (Table 1).

The author's method was to list the ca. 5,400 different words of the Greek New Testament—grammatical variants of the same vocabulary item being regarded as the same word—and to enter for each word its frequency of occurrence in each of the books of the N.T.[2] Summation of these numbers then gives the frequency of a particular word in the whole N.T. On the other hand, summation of words having the same frequency gives the frequency distribution of vocabulary according to frequency of occurrence, for each part of the N.T. and for the whole.

Table 2 gives the frequency distribution of vocabulary for the whole New Testament and its parts (apart from the Catholic letters) in abridged form.

From Morgenthaler's basic table (91 pages) we can obtain not only the frequency distribution for the whole N.T. and its parts, but also the number of words peculiar to one book, two books, etc. As will be seen in Chpts. III and XVIII, this provides the material for tackling the important problem of vocabulary partitioning.

[2] For New Testament.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 2

Occurrence number	Word number (Vocabulary)								
	New Testament	Matthew	Mark	Luke	John	Acts	Paul	Hebrews	Revelation
1	1934	670	634	971	375	943	1140	575	310
2	842	276	219	352	153	335	442	166	165
3	470	169	108	170	92	182	230	78	97
4	322	89	71	107	64	94	143	54	50
5	221	70	52	55	36	82	123	26	37
6	173	54	36	48	27	41	62	26	32
7	123	59	18	43	16	36	58	9	28
8	110	30	23	23	20	43	50	10	32
9	107	27	16	27	16	22	35	11	16
10	79	22	15	23	8	25	30	10	14
11	83	17	12	16	14	23	17	5	8
12	64	16	16	15	13	12	22	4	10
13	42	17	5	19	14	14	15	7	13
14	52	8	3	13	8	11	27	4	6
15	42	8	12	14	11	11	15	2	5
16	39	9	3	9	6	14	15	5	5
17	33	9	5	14	15	10	14	4	4
18	36	7	7	8	8	5	11	1	5
19	19	10	6	7	5	10	8	5	6
20	18	5	7	6	3	7	7	2	3
21	19	5	5	4	5	9	10	3	6
22	21	6	4	3	1	7	5	1	2
23	20	1	4	3	4	5	8	3	7
24	14	3	4	1	3	4	5	1	2
25	17	5	2	6	3	6	2	1	3
26	17	4	2	6	3	2	6		4
27	14	4	2	4	2	2	8		2
28	14	4	1	4	4	2	3	1	1
29	17	5		4	1	2	5		3
30	9	2	2	1	3	3	4	2	1
31	16	2		5	3	1	1	2	
32	10	2	1	1	2	1	2	2	
33	5	1	3	5	3	1	4		1
34	8	1	2	2	3	1	5		1
35	7	1	1	1		5	2	1	1
36	10	2	2	2	5	2	2		1
37	9	3		1		2	2		1
38	6		1	1			2		2
39	7	1	2	1	1		2	1	
40	6	2			4		5		

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Table 2, continued

Occurrence number	Word number (Vocabulary)								
	New Testament	Matthew	Mark	Luke	John	Acts	Paul	Hebrews	Revelation
41	11	2		2		1	2	2	
42	11	2	2			1	2		
43	4	2	1	2	1	1	4		
44	7		1	2		1	1	1	1
45	7	2		3		2			
46	4		1	3	1	2			1
47	9	1	2			1	4		1
48	5	1	1		1	1	1		
49	5	2			1	1	1		1
50	6	1		1	1		2		
1–50	5124	1639	1314	2008	959	1986	2564	1025	888
50+	312	52	31	47	52	52	84	13	28
Total	5436	1691	1345	2055	1011	2038	2648	1038	916

As is well known, vocabulary statistics have been, and are, used with great advantage in the exegesis of the New Testament and the Septuagint (see the statistical investigations by writers like J. Ritchie Smith, H. J. Cadbury, A. v. Harnack, P. N. Harrison, E. Percy, and specially J. C. Hawkins' *Horae Synopticae* as the most extensive work of this kind). What is new in the science of statistical linguistics is not only the improvement of the sampling methods for word counts and their evaluation by the rules of statistical inference, but the tendency to derive from them basic information about language structure. Investigations of disputed authorship, chronology, etc. by means of statistics, though very useful for their particular purpose, are of secondary importance compared with research on problems of structural linguistics.

Stylostatistics. The basis for the quantitative characterisation of style is the frequency distribution of vocabulary according to frequency of occurrence of individual vocabulary items. The characterisation, as a basis for comparison, would conceivably be in terms of the relation between vocabulary and occurrence (text length in terms of number of words), and in terms of either, that is of vocabulary alone or of frequency of occurrence alone.

The first type of characterisation may again be in terms of the whole distribution or in terms of an overall statistical parameter, such as the Type/Token ratio, the logarithmic Type/Token ratio and the Characteristic $K(v_m)$; the second may be in terms of the partitioning of vocabulary among the parts of a whole or in terms of probability of occurrence of particular vocabulary items.

The various methods of quantitative style characterisation are dealt with in Chpts. I, II, III, IV, V.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

1.4 THE LOGARITHMIC TYPE/TOKEN RATIO

1.4.1 Previous observations on the logarithmic Type/Token ratio

By the term "type/token ratio" we understand the ratio of the vocabulary in a given text, or sample from it, to the total number of words comprised in the text or the sample. That quantity changes, in general, with the size of the piece of literary work, the vocabulary increasing with the text length—but by no means proportional to it—in such a way that the quantity decreases, on the whole, with increasing sample size. It cannot, therefore, in this form serve as a characteristic of style, which must be independent of the text length. The logarithmic type/token ratio, i.e. the log type/log token, on the other hand, remains sensibly constant for samples of different size from a given literary text and is, therefore, suitable to serve as a style characteristic.

This fact—one of the most remarkable in the field of quantitative linguistics—appears to have been independently discovered by three investigators: Chotlos (1944), Herdan (1957) and Devooght (1957). Chotlos used it as an alternative to another relation which had been designed in accordance with, and as a corollary of, the Zipf law, and which Chotlos did not find to fit the facts. For Chotlos the relation was an empirical one and he offered no explanation for the observed regularity.

The present author, in a paper read before the 8th International Congress of Linguists in Oslo (1957) showed that the log vocabulary plotted against the log text length of 61 works by Chaucer showed, by and large, a straight line trend, and that such a relation could be derived under the assumption that the 61 works by Chaucer were, as regards vocabulary size or the type/token relation, random samples withdrawn from the frequency distribution of the total vocabulary at Chaucer's disposal.

Lastly, Devooght has derived the relation in question from the modification of the Zipf law known as the Mandelbrot Canonical law. It appears, however, that to use the Zipf-Mandelbrot law as the starting point for the derivation of the bilogarithmic type/token relation makes matters unnecessarily complicated, and even unsatisfactory from a mathematical point of view, since, in order to arrive at the relation, Devooght has to make certain assumptions which are not in accordance with reality, e.g. to let the vocabulary approach infinity.

We shall give here another derivation of the relationship in question as a general linguistic law which is free from both the complexity and the shortcomings of Devooght's derivation. This is made possible by the fact—overlooked so far—that the relation in question is what is known in biology as the law of relative growth.

The problem of relative, or as it is also called, differential growth is a fundamental one for biology. All organic forms, save the simplest such as the spherical or the amoeboid, are the result of differential growth. Insofar as growth relations are subject to the law of relative growth, they are capable of quantitative formulation. In general, we can say that the relative growth of certain organs stands in a constant relation to that of the whole organism. If x is the magnitude of the animal (in terms of some standard linear unit of measurement or in some unit of weight) and y the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

magnitude of the differentially growing organ, then the relation between them is of the form of a power function

$$y = bx^{\gamma} \tag{1}$$

where b and γ are constants. The constant b is here of no particular biological significance, since it merely denotes the value of y when $x = 1$, but the value of γ has an important meaning. It implies, as we shall see, that, for the range over which the formula holds, the ratio of the relative growth rate of the organ to the relative growth rate of the body remains constant, that ratio itself being the value of γ.

1.4.2. The law of relative growth of literary products

The law can be deduced on the basis of very simple assumptions about growth in general. One essential fact about growth is that the rate of growth of a system growing equally in all its parts is at any moment proportional to the size of the system. A second fundamental fact about growth is that the rate of growth slows down with increasing age, i.e. size of the system; a third is that it is affected by the external environment.

Whereas the rate of growth of particular organs may differ from that of the whole system, the two latter considerations affect all parts of the system equally, so that we may suppose that the growth-rate of any particular organ is proportional simultaneously (a) to a specific constant characteristic of the organ in question, (b) to the size of the organ at any instant, and (c) to a general factor dependent upon age and environment which is the same for all parts of the body.

We now make use of these ideas for setting up the differential equation for the growth of vocabulary with that of text length in literary works. If the system is one of linguistic material, we may suppose that the growth rate of the vocabulary is proportional simultaneously (a) to a specific constant characteristic of the accumulation of vocabulary with increasing text length, (b) to the magnitude of vocabulary that has already accumulated, and (c) to the particular conditions of writing as due to style, content etc. The same applies to text length, as a characteristic of the whole system.

If V stands for the number of vocabulary items at the time t of composition, and N for the text length, we shall then have

$$dV/dt = \beta VG$$

and

$$dN/dt = \alpha NG \tag{2}$$

where α and β are the specific constants for the text length and the vocabulary, and G measures the "environmental" conditions of growth of linguistic systems as affected by the style of the writer and the content of the text. Eliminating G we get

$$dV/dN = \beta V/\alpha N$$

and

$$dV/V = \frac{\beta}{\alpha} dN/N \tag{3}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Integration gives

$$\log V = \frac{\beta}{\alpha} \log N + \log b$$

or

$$V = bN^{\beta/\alpha} \tag{4}$$

where b is a constant. We know that it denotes the value of V when $N = 1$; but a sentence containing one word must also contain one vocabulary item, and can have not more than one vocabulary item. It follows that for our distributions b is unity, and $\log b = 0$, and we have, writing γ for the constant ratio β/α

$$\log V = \gamma \log N$$

or

$$\log V/\log N = \gamma \tag{5}$$

The relation of relative growth of organs can be established either by following up the growth of an individual or of a group of individuals belonging to the same species (under similar environmental conditions). As an example, we may take the relative growth of the antlers of deer. As is well known, these are shed each year, and replaced the year after by totally new growths. Usually each new growth is larger than the preceding growth, the proportional increase in antler-weight being a linear function of the proportional increase in body-weight during that year. Instead of establishing this relation by following up the growth of an individual deer, it can be established for a great number of deer under the same environmental conditions by relating the average antler-weight to the average body-weight of animals of different age. Figure 1 shows the bilogarithmic straight line relation which exists between antler-weight and body-weight in 527 adult red deer (*Cervus elaphus*).

This implies that the relation of relative growth is not restricted to concomitant changes of parts or aspects of an individual organism, but may be extended to such changes in different individuals belonging to a homogeneous group. In its application to language this means that the constancy of the bilogarithmic type/token ratio is characteristic not only for the parts of a given text, but for a number of texts which could be regarded as parts of a homogeneous whole (system).

This is an important extension of the stability of the $\log V/\log N$ ratio as conceived by either Chotlos or Devooght, according to whom it referred only to samples of different size of one and the same text. In our derivation there is nothing which implies the need for such a restriction. As long as the "environmental" conditions symbolised by G remain the same—and they may do so not only "within" a text but from text to text—the ratio $\log V/\log N$ is constant.

This extension renders the bilogarithmic type/token ratio a valuable criterion of style, which could be used in addition to, or in place of, other quantitative style criteria.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

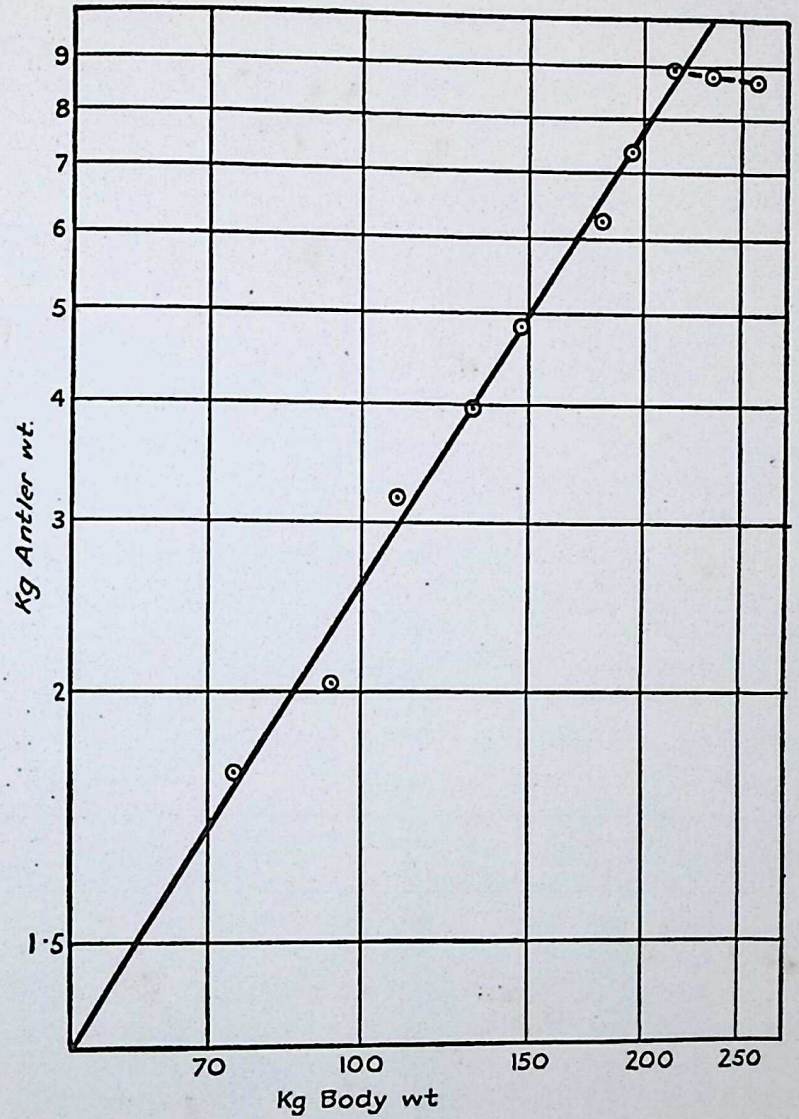


FIG. 1. Antler-weight against body-weight in 527 adult red deer (*Cervus elaphus*); logarithmic plotting (Huxley, 1932).

1.4.3. The constant differential growth-rate as a style characteristic

As test cases we have chosen two parts of the New Testament in the original Greek: the Pastoral Letters (1 Timothy, 2 Timothy, Titus) on the one hand, and the Gospel according to St. John, the letters of St. John among the Catholic letters, and the Revelation, on the other.

The authorship of the Pastorals must have been among the first instances inviting the application of quantitative methods in literature. Schleiermacher's critical

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

campaign against the authenticity of Timothy 1 uses a great array of hapax legomena for establishing the linguistic peculiarity of the Epistle. The controversy continued on these lines, subsequent workers using linguistic criteria to a greater or lesser extent according to the value they attributed to them. This continued until 1921 when Harrison's book *The Problem of the Pastoral Epistles* appeared, in which he presented such overwhelming linguistic evidence against Paul's authorship that most workers considered the controversy closed. However, in 1957, D. Guthrie delivered his attack on Harrison's conclusions in his book *The Pastoral Epistles*, and the problem is still regarded as unsolved (Grayston and Herdan, 1959).

The second group of texts mentioned above comprises the writings nominally due to St. John, and one work for which John's authorship is sometimes assumed: the Revelation. Textual criticism has isolated these writings as having certain characteristics which distinguish them from the rest of the New Testament, such as a small vocabulary, few composite words, and few words occurring in the Septuaginta. They are regarded as being the work of Palestinian Hellenists.

The following tables list the numerical characteristics for the series of the Pauline Epistles on the one hand, and the books of the New Testament on the other, against which those of the particular parts in which we are interested must be judged. The material is taken from R. Morgenthaler's *Statistik des Neutestamentlichen Wortschatzes* (1957) which represents a thoroughly reliable statistical exposition of New Testament Greek.

TABLE 3

Pauline Epistle	Text Length (L)	Vocabulary (V)	Type/Token ratio (V/N)	γ ($\log V/\log N$)
Ro.	7094	1068	.1505	.7864
1 Co.	6807	967	.1420	.7788
2 Co.	4448	792	.1780	.7944
Gal.	2220	526	.2369	.8131
Eph.	2425	529	.2181	.8046
Phi.	1624	448	.2758	.8257
Col.	1577	431	.2733	.8238
1 The	1472	366	.2486	.8092
2 The	824	250	.3033	.8223
1 Ti. }	1586	541	.3411	.8540
2 Ti. } Pst.	1235	458	.3705	.8605
Tit. }	663	303	.4570	.8794
Phe.	328	141	.4298	.8542
All Paulines	32303	2648	.0819	.7590

a) *The Pauline Epistles (Greek)*

Table 3 gives in cols. 2 and 3 the text length and vocabulary size for the 13 Pauline Epistles in the original Greek, and in col. 5 the values for γ for the individual Epistles

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

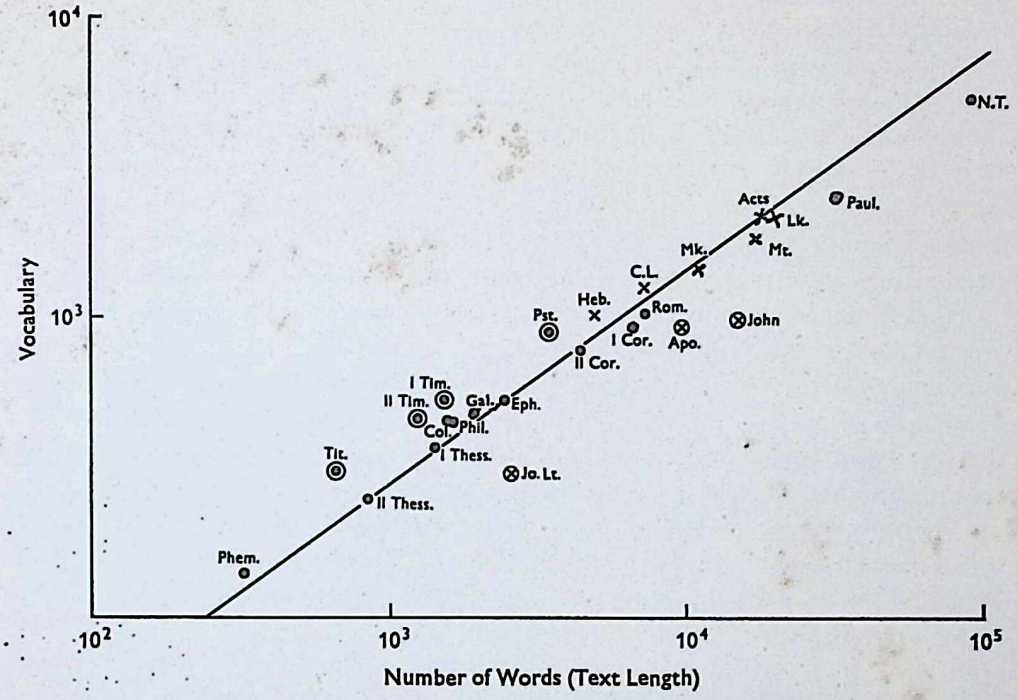


FIG. 2. The logarithmic Type-Token ratio

TABLE 4

New Testament Book	Text length (L)	Vocabulary (V)	Type/Token ratio (V/N)	γ (log V/log N)
Mt.	18278	1691	.092	.7573
Mk.	11229	1345	.120	.7724
Lk.	19404	2055	.106	.7726
Jh.	15420	1011	.066	.7174
Acts	18374	2038	.111	.7760
Pls.	32303	2648	.082	.7590
Hb.	4942	1038	.210	.8165
Catholic letter	7560	1271	.168	.8003
Letters of John	2597	302	.116	.7263
Apo.	9818	916	.093	.7324
Whole N.T.	137328 (89883)[3]	5436 (5427)	.060	.7535

Average log type/token ratio (leaving out Gospel according to St. John, letters of John and Revelation) results as .78 approximately.

[3] The excessive frequency of grammar forms, e.g. over 19,000 for the articles, made it advisable for the purpose of calculating the type/token ratio to exclude from both vocabulary and text length words with a frequency of occurrence exceeding 1700 (11 words having 47,544 occurrences).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

and also for the total. Apart from the 3 Pastorals (1 Ti., 2 Ti., Tit.), γ can be regarded as sensibly constant, and we may regard its average for the 10 Pauline Epistles (that is, without the Pastorals), $\gamma = .8113$, as epitomising the logarithmic type/token relation for the Pauline Epistles in general. Formula (5) thus becomes for that series

$$\log V/\log N = .8113$$

For the purpose of comparison, we have given in col. 4 the conventional type/token relation for each letter and for the whole which can be seen to be anything but constant.

The plot of these data on a bilogarithmic grid shows a very satisfactory straight line relation between log V and log N for the individual Pauline Epistles and the total of vocabulary and text length in all Paulines. This is as it should be since any individual Pauline Epistle can be regarded as a sample from the total. It is interesting to note that the Pastorals (cumulatively and individually according to the 3 Epistles they comprise) are the proverbial exception to the rule, the points being significantly above the general trend of log V and log N for the Paulines. This must be regarded as linguistic evidence for a characteristic difference in style between the Pastorals and the remaining Pauline letters and thus lends support to the theory of non-Pauline authorship for these Epistles.

b) *The Books of the New Testament (Greek)*

Apart from the points (.) representing the Paulines, the graph shows also crosses (×) for the other books of the New Testament and the New Testament as a whole. It is now one of the most striking facts in support of the hypothesis of a constant differential growth-rate throughout the whole of the Greek New Testament that both sets of points show the same linear trend which means, according to formula (5), a constant differential growth-rate throughout the New Testament of about .80, the Pastorals above the trend line and the group of John's writings together with the Revelation below that line being the proverbial exceptions which confirm the rule. Inspection of table 4 shows that the differential growth-rates for the last-named group of writings are distinctly different from that of the other parts of the New Testament, and also from that of the whole, but are very similar among themselves which supports the hypothesis of a common author for the writings nominally attributed to John and the Revelation.

1.4.4. Discussion of method

Among many linguists there still appears to linger a distrust of the application of even such elementary mathematics to language problems. The usual criticism is that the formulae arrived at may have a certain convenience, but tell us nothing new and nothing worth knowing about the linguistic side of the phenomenon. This appears to be very ill-founded.

In the first place, to have a quantitative expression in place of a vague idea of a general tendency of vocabulary to grow with text length is not only a mild convenience. Let us recall that a constant differential growth-rate for literary products is not merely

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

an empirical rule found over a large range of literary texts, but represents the conclusion reached by considering production of literary work as complying with the general characteristics of growth in the organic world, and thus something to be expected on theoretical grounds.

Moreover, it was shown that it has practical uses, since it may be applied when we are called upon to make certain deductions about similarity and dissimilarity of style. Using it in that capacity we found that it led to the same conclusion which had been reached independently, though tentatively only, by the conventional methods of textual criticism, and thus appears to be eminently significant linguistically. In this capacity it may be found useful in the investigation of texts, like the Dead Sea Scrolls, whose authorship is problematic. Among the Qumran fragments there are also some in Greek. One might expect the language of these fragments to show the same statistical characteristics as those parts of the New Testament which are the work of Palestinian Hellenists, to whom we today ascribe such works as the Gospel according to St. John, the Letters of John, and the Revelation.

1.5 THE LINGUISTIC FREQUENCY DISTRIBUTION

The concept of the linguistic frequency distribution is not identical with the so-called Zipf law, although the latter would seem to spring to the mind of linguists as well as mathematicians whenever "frequency" is mentioned. The Zipf law is a particular form of frequency distribution claimed to result when frequency of occurrence of words—different grammar forms of the same word being counted as different words—is plotted against the "rank" of these words, the word with the highest frequency being accorded the lowest rank 1, and the subsequent words the ranks 2, 3 . . .

It is difficult to understand why the Zipf law should have attained such notoriety, if not fame, since it is not, to the writer's knowledge, of much practical use to the linguist, and mathematically a triviality.

The Zipf law is the bilogarithmic straight line relation between the frequency of occurrence of words in a language and their rank. Mathematicians believe in it because they think that linguists have established it to be a linguistic law, and linguists believe in it because they, on their part, think that mathematicians have established it to be a mathematical law. As can be shown in a few lines, it is not a law at all, in the sense in which we speak of natural laws. I do not here refer to the question of whether the straight line relation was a good fit of the observed data. Doubts in this respect have been voiced by linguists (Guiraud, 1955) and mathematicians (Mandelbrot). Guiraud's excellent critical examination of the Zipf law leads to a virtual replacement of that hypothesis by an empirical formula. But nobody, to the writer's knowledge, has yet questioned the correctness of the term "law" which had been accorded to it. And yet, it is no law.

If the words in the sample under consideration are arranged in descending order, say, with regard to frequency and the other variable is arbitrarily made a sort of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

inverse function of frequency by giving the word with the highest frequency the lowest rank, continuing with this along the series of the integral numbers, then it follows as night follows day that the relation between our variables must be an inverse one. Having made the second variable an inverse function of the first, we surely cannot be surprised to find the relation an inverse one. It is true, the relation need not have turned out to approach the bilogarithmic straight line relation. But that it happens to do so does not make it any more a law. It remains the consequence of the arbitrary nature of the variable "rank". This does not mean that we may not find a truly linguistic variable which substituted for "rank" would lead to a relation of similar form.

It may be advisable to emphasise the difference between statistical linguistics as understood by the present author and the introduction of mathematical methods into linguistics by Zipf.

What Zipf was trying to do when setting up formulae for language relations was to establish *functional relations*[4] in linguistics. That is to say, he tried to see mathematical or physical laws as applying to language, e.g. the law of least effort.

This is very—if not fundamentally—different from treating language material like other empirical material which, for want of clear-cut functional relations between its elements, is observed with a view to establishing at least *statistical regularities* among its events. It is true that Zipf did some work also on these lines in his studies on the frequency distribution of linguistic forms, but even there he tried to set up functional —not statistical—laws, such as the one connecting the number of vocabulary items with their frequency of occurrence. To the author's knowledge, he never tackled the very basic problem of the *possibility* of applying statistical methods to language. The justification for speaking of language in terms of mathematical statistics was given in Chapter 5 of Volume I, where it was shown that de Saussure's dichotomy of language into "langue" and "parole" can be made to correspond to that between the statistical population and the sample by letting "langue" stand for the totality of engrams in the brains of members of the speech community together with their probabilities of occurrence, and "parole" for random samples from it.

Since then a great step forward was made through the discovery that the basic probabilities are what linguists call the functional burdening of linguistic qualities. Originally the term referred to phonemes, but we shall use it as a general term for the distribution in the dictionary of any linguistic characteristic as against the frequency distribution of such characteristics in speech or writing. In Chpt. VIII the relation between the functional burdening of phonemes and their frequency of use in speech is shown to be of the nature of the relation between the statistical population of events at risk, and their occurrence through random sampling, and in Chpt. IX the relation between the distribution of word length in "la parole" is shown to follow from that in the dictionary, or as we might say, the utilisation of word length is a consequence of its functional burdening.

[4] In the mathematical sense.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

For vocabulary items, or briefly for meaning, a dictionary count is possible, but fraught with great difficulties. So far no really satisfactory numerical investigation has been carried out to show that the occurrence distribution of vocabulary items in "la parole" is a random sample of their functional burdening with meaning. However, a tentative statement to that effect with some numerical material to support it was made in Sect. 6.1 and Sect. 11.5, Vol. 1, and Chpt. II of this volume affords strong support for it.

This development means that the former hypothesis of "la langue" as comprising the vocabulary (in its widest sense) as dictionary material together with the probabilities of occurrence is now replaced by a scientific theory in which those probabilities are empirically ascertainable properties different from the corresponding frequencies in "la parole". They are the properties of categories of linguistic form (simple or complex) in the vocabulary (or, roughly, dictionary) at risk, as different from their properties of occurrence in speech or writing.

1.6 A CRITICAL EXAMINATION OF THE ZIPF LAW

The Zipf law is the product of a period when quantitative methods were a novelty in linguistics. What was an achievement then is quite obsolete now. Zipf's conception of statistics is very primitive, judged by modern standards, and so are his ideas about linguistic laws as replicas of physical laws, e.g. the explanation of linguistic behaviour by the "law of least effort". With the mathematical tools which are now at our disposal, it is almost an anachronism to take the Zipf law as a sort of basic and incontrovertible truth. We hereby only create difficulties for ourselves, for it is not a law in the theoretical sense (1), and it is not empirically true (2). "Our science has reached a state where rigor is demanded in all its aspects" (W. A. Austin, *Language*, 33, 1957, p. 538).

To (1): It has been shown above that the bilogarithmic straight line relation between the occurrence frequency of a word and its rank cannot be regarded as a scientific law in the correct sense of the term. That the decrease of frequency should be related to an increase in rank follows not from any natural property of language structure, but merely from the fact that the word with the highest frequency is given the lowest rank, and as the frequency decreases the words are given correspondingly higher ranks. Thus the inverse relation between frequency and rank which is at the basis of the so-called Zipf law is one of our own making.

To (2): Quite apart from the fact that Zipf counted every occurrence as a word, that is every paradigmatic variant of a vocabulary item as a word by itself, whereas some of his followers count all grammatical variants of a vocabulary item as one word, all kinds of exceptions have had to be suggested to make the "law" fit the actual observations. According to some investigators, it does not hold for high-frequency words, nor does it hold for the low-frequency words, but seems to fit only the distribution of words of intermediate frequency. Considering that no definition is given

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

— nor can it be given — for high- and low-frequency, and consequently for the "intermediate" frequency for which the "law" is claimed to hold, it is evident that we cannot speak here of a law, if we use the term in the sense in which it is used in science today.

Not only are there all kinds of ill-defined "exceptions" to the "law", but various hypotheses of doubtful correctness have been introduced to justify the Zipf law, e.g. that by Joos (see below). Of another kind is Mandelbrot's derivation of the Zipf law. He substitutes for the quasi-variable "rank" a hypothetical variable, which he defines as "cost of transmission of information", and arrives at a function which, under certain conditions, goes over into a bilogarithmic straight line similar to the Zipf law.

How true it is that by adhering to the Zipf law in the face of so much evidence against it we only create difficulties for ourselves, can be seen from its use for estimating the size of a writer's vocabulary.

By "estimation of vocabulary" we may understand very different things. It may refer to the hypothetical total of words in the mind of a writer upon which he draws as he works; in this sense the phrase was used by Yule and is used by Guiraud, also by Good (1953). Since that total of words at the disposal of a writer is a rather nebulous concept—there being little or no agreement between the way of estimation suggested by the three writers—, we turn to the second interpretation of the phrase, which refers to the total of different words in a given text or composite sample of texts (also of spoken language, of course) as against the vocabulary of parts or samples from it. This is what we shall mostly mean by "estimation of vocabulary".

According to Zipf, the product of the relative frequency of occurrence of a word F and its rank R is a constant C. Zipf now claimed that the products $F \times R$ were all of the order 1/10, and that the law therefore assumes the form $F = 1/10R$, and this is repeated like a Gospel truth over and again in the literature, by linguists and scientists alike (Brillouin), although examination on even a limited scale would have shown that it is simply not true. If it were, it would lead to the absurd conclusion that every word sample contained 12,000 different words regardless of length, since it takes about 12,000 terms of the series $1/10 + 1/20 + 1/30 \ldots$ to equal 1. It follows that according to the Zipf law, every sample examined must contain the same vocabulary, 12,000 words![5]

To overcome these contradictions, much doctoring is needed if one still sticks to the "law", and one of these is the modification by Joos who takes the probability (instead of the relative frequency) of a word of rank R to be

$$P(R) = 1/10R^{\alpha}$$

Using now the modified version of the "law" as proposed by Joos under the name

[5] $1/10 + 1/20 + \ldots + 1/10R = 1/10\,(\log_e R_{max} + \gamma) = 1$ where γ is Euler's constant
$\log_e R_{max} = 10 - \gamma = 9.4$
$R_{max} = e^{9.4} = 12{,}000.$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of a "probabilistic model", the vocabulary size in a sample of n occurrences is calculated as

$$V(n) = \Sigma\{1 - [1 - P(R)]^n\}$$

or

$$V(n) = \Sigma\{1 - (1 - 1/10R^a)^n\} \quad (6)$$

This formula was used by A. Koutsoudas (*Language*, 33, 1957, p. 544-552) for calculating the vocabulary size of samples from Pushkin's *The Captain's Daughter*. He used first a hypothetical vocabulary of infinite size (whatever sense there is in this) and then one of 600,000 words, only to find that the estimation was in no way near the actual vocabulary in the novel. Then—without realising that this meant changing the conception of a hypothetical to an experimental vocabulary—he estimated the vocabulary, M, from the experimental values of the vocabulary in samples of 10,000, 20,000, and 30,000 words of the novel, which if introduced in formula (6) give $M = 5{,}530$. Now this estimation of vocabulary is more in accordance with the facts, though still not very satisfactorily so, since the agreement between observation and theory is not very good. The lesson Koutsoudas draws from this is that the coefficient 10 in the denominator of the Zipf formula was not correct.

Thus, to summarise, all evidence goes to show that not only is the "law" invalid for high and low frequencies, but the reciprocity of F and R has to give way to one of F and a power of R, and finally, the coefficient 10 in the denominator is not correct. Moreover, even if the form of the "law" may be said to have some generality, both constants, the coefficient and the power of R have to be determined anew for every new text. What is then left of the Zipf law? Whatever it is, the relation should no longer be referred to as a law, since this can only lead to confusion, the simple and straightforward relation between vocabulary and occurrence which it suggests just not being true.

The modifications of the original Zipf Law by Mandelbrot (1956),[6] Belevitch (1957)[7] and others, trying to substitute for "rank" a truly linguistic variable have, in my opinion, not succeeded. Both Mandelbrot and Belevitch substitute first "cost of transmission" (i.e. by telecommunication devices) which is then simply equated to "word length".[7] A brief consideration of the actual linguistic facts involved would have shown that this meant abandoning the double-logarithmic law.

The scale of "rank" is theoretically infinite, and even for moderately large texts extends to over 10,000. (See the bilogarithmic graphs in Mandelbrot.)

Word-length, on the other hand, even if expressed in terms of the ultimate linguistic units of the spoken or written language—phonemes or alphabetic symbols—does not exceed, at any rate for our Western languages, 12-15 units, and, in fact, tails off very quickly after a peak at, say, 3-5 units. The discrepancy in dispersion between the two variables, frequency and word-length, makes it impossible that their relation should be satisfactorily fitted by the bilogarithmic straight line law.

[6] B. Mandelbrot, et al., *Logique, langage et théorie de l'information.* (Paris, 1957).
[7] V. Belevitch, *Langage des Machines et langage humain* (Bruxelles, 1956).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Since each length-category comprises a number of items, all of the same length—such numbers being quite appreciable for the lower length categories—but with different frequencies, the co-ordinate representation of all the items belonging to one category will be a straight line vertical to the word-length axis, and instead of a straight line graph we would get a rather pronounced step-wise descent of the line.

Moreover, it is established practice of modern scientific procedure that if a mathematical model is used for the explanation of an observed relation—as in this case—, the "goodness of fit" is tested by an appropriate test function, say Chi-square. Yet nowhere in the relevant literature have I come across such a test carried out on the modified Zipf-Mandelbrot law with "cost" or word-length as the dependent variable.

However, the Zipf law, although unsuitable for the scientific description of linguistic distributions, has its uses when it comes to the mechanical handling of word masses. (See Part III, Sect. 15.2.2). This does not involve a contradiction. Apart from the fact that the bilogarithmic straight line relation fits part of the distribution—the medium range of rank—and insofar can be said to be true, the aims of the statistical description and that of mechanical handling of data are very different from one another, and it is quite feasible that what is unsuitable for one should be useful for the other.

The Zipf law cannot claim to be regarded, strictly speaking, as a scientific law, since it does not satisfy the criteria for such laws, yet it is often sufficiently close to the actual distribution to be of service in the technology of language, and we may regard it as a useful technological device.

It needs no emphasising that only the ambivalent consideration of language structure, from the viewpoint of statistics, and from that of mechanisation, can be regarded as complete. To declare one as irrelevant because one happens to be by profession concerned with the other, shows lack of understanding.[8]

1.7 IN DEFENCE OF THE "STATISTICAL HYPOTHESIS" FOR THE USE OF WORDS

It is advisable at the outset to clear up one of the most frequent misunderstandings of what statistical methods of this type imply. The misunderstanding in question stems from the somewhat unfortunate analogy used by Yule when describing the statistical hypothesis of word occurrence.

"The scheme that gives rise to the word distribution may be pictured somewhat as follows. Let us suppose all the words in the writer's treasure-chest to be written out on tickets, some—the words most rarely used—on only one ticket each, some on two tickets, some on three, and so on; some few words very frequently employed, on several hundreds or even thousands of tickets each. We regard the process of writing (it does sound rather ridiculous put in this way) as bearing an analogy to the drawing of tickets out of this chest" (Yule, 1944).

[8] For a one-sided criticism of this type by an exponent of mechanical translation, see A. D. Booth in *Nature*, vol. 179 (1957), p. 837.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

There are certain rather typical objections against attaching importance to the characteristic of relative frequency of words or, more precisely, against the statistical hypothesis of the stability of relative frequencies of language forms as a characteristic language feature.

The determinist linguist insists that if he uses a particular word, say, "teapot", it is because he wants to say just "teapot" and not "kettle" (Ross, 1952), and he believes he has demolished the statistical hypothesis, like Dr. Johnson, when he disposed of Berkeley's idealistic philosophy by kicking a stone: "I refute it thus, sir."

This criticism is not to the point from whatever angle we would like to look at it. From the linguistic angle it is not to the point, because a single word like "teapot" cannot be described as language in the sense in which de Saussure uses the term, and according to whom a word has its meaning not as an independent unit but only in conjunction with all the other words of the language. From the statistical angle it is not to the point because statistics does not deal with single events as such, but only with masses of events, and with a single event only against the background of the mass event. What the statistical hypothesis which we use in our work implies is this: if from a given text samples of word occurrences are withdrawn from different parts, each comprising a sufficient number of words, then the frequency distribution of vocabulary items according to their frequency of occurrence will not be significantly different from sample to sample, and from sample to total text. In other words, we may regard any part of the text of sufficient size to be virtually a random sample of the total text as regards the frequency distribution of vocabulary.

Thus, the statistical hypothesis of word occurrence does not purport to explain *why* a particular word is used, nor why it is used in a particular grammatical combination. On the contrary, it shows that no matter what particular words are used in a given context, and no matter how they are put together according to the rules of grammar, a sufficiently large part of the text can be regarded as a random sample with respect to word frequency (and, as was shown in Vol. I, also with respect to other characteristics such as phonemes, word length in terms of letter, phoneme, syllable number) or, more precisely, as a random sample of the population probabilities of the use of words (phonemes, letters, word length etc.) in "la langue".

That the use of words in speech and writing is primarily a matter of choice, nobody will deny. But this does not preclude certain linguistic characteristics of those words from being what we call "statistical or chance variables", that is, variables which take on different values according to one or the other law of chance. On the contrary: it follows from the principle, enunciated by de Saussure, of the independence of word form (signifiant) from word content (signifié).

Speaking (writing) may be regarded as selecting a sample of word material according to the speaker's (writer's) individual *choice*. Since the linguistic properties, such as phonemic structure, length, etc. are independent of the meaning of the words, and with it of their particular choice, it follows from a theorem of mathematical statistics that the masses of those linguistic characteristics in the sample will represent a

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

random sample of the respective statistical populations of those linguistic properties. This represents the element of *chance* in the use of words.

The statistical properties of the linguistic elements are the field of language which is amenable to statistical treatment. That field comprises among other items: the frequency distributions of letters and phonemes, of word length in terms of letter-, phoneme-, syllable-number, the repeat rate of vocabulary items and other linguistic forms, the frequency distribution of vocabulary, the vocabulary connectivity between the parts of a text and between texts forming a more or less homogeneous group.

But hardly have we disposed of the objection to statistical methods on the grounds that chance has no place in language, than we have to face the opposite objection, also raised rather unexpectedly by linguists,[9] that insofar as chance *does* play a role in the use of linguistic forms, their occurrence just obeyed certain general laws of nature, and stylostatistics was, therefore, linguistically irrelevant. To be more precise: it is sometimes objected that statistical uniformities of language and the statistical laws summarising such uniformities, are of little *linguistic* interest because they simply show language as being subject to such laws as have universal application whenever events are the result of a multiplicity of a great number of small causes.

Apart from the fact that it is deserving in itself to establish language as being subject to such laws—a thing which is by no means a priori certain—, so sweeping a view misses the important point that it is precisely the statistical nature of certain linguistic variables which provides the basis for what we call stylostatistics, which is eminently useful in language research. As we ascend from the smallest units—letters, phonemes—to larger ones—morphemes, syllables, words, sentences—, subject (content) and individual style begin to exert a more and more pronounced influence upon the statistical variable, and differences in the statistical parameters may become significant of changes in both, content and style. This is the reason why statistical linguistics, as the basis of stylostatistics, is so important a tool in literary research.

Another type of objection is concerned with grammar and may be epitomised by the following criticism of the statistical hypothesis and, in particular, of Yule's treasure-chest analogy. "Such a model is fundamentally a negation of syntax, for it does not distinguish a normal English book from one in which the identical set of words occurs in inverse order, or one in which the same words were scrambled up in any one of a multitude of conceivable ways" (Halle, 1958).

What is meant by this is evidently that a given frequency distribution of words does not uniquely characterise a text, since the grammatical combinations in which the words occur are not specified; in other words, the content is not uniquely determined. With this, however, the critic has not brought to light something which has been overlooked by the statistician. On the contrary, the statistician is fully aware of the fact that the frequency distribution characterises style, that is *form*, not *content*. That is why Yule's K is used as a style characteristic, not as a content characteristic. But even so, no uniqueness in the strict mathematical sense is implied, only in the statistical

[9] N. Chomsky, *Language*, 34 (1958), p. 100.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

sense. It should be realised—and this is of great importance—that in statistics we never prove, but only aim at stating the probability with which an event may occur. If that probability is very small, it may, for practical purposes, be neglected. What the logician therefore ought to ask himself is *how likely it is that of two books containing the identical words with the identical frequencies, and therefore being statistically similar or even identical, one was written back foremost.* Since the probability of such an event is truly negligible, we may disregard this type of objection to statistical linguistics.

Regarding the equally typical reproach that what has been ascertained in the way of statistical regularities has no linguistic relevance, we ask: is it not linguistically relevant that no matter how carefully and individually words are supposed to be chosen and grammatically connected in speech and writing, there should be something (and a great deal, at that) which is beyond the intentional or conscious control of the speaker (writer)? The proof that this is "linguistically interesting" as containing new and relevant information about language lies in the very fact that some linguists deny that there *is* such a stability of the quantitative elements in language, and sometimes even argue that it *could not be*!

It cannot be said too often: statistical linguistics is basically an empirical science and its results should be examined with regard to whether they are true, not whether in some way or other they seem to contradict the subjective impression of the linguist. An almost perfect example of misinterpretation of statistical linguistics is provided by the contribution by H. Glinz to the discussion on "Mathematical Linguistics", p. 82 of the *Proceedings of the 8th International Congress of Linguists.*

He took the 220 most frequent German words according to the Frequency Dictionary by Kaeding, and sorted them into categories: verbs, nouns, adjectives, pronouns, particles. As regards particles (75), pronouns (35) and verbs (28), he claims to have found that the most frequent words were also the most general and best known ones, and that they show what he calls an "inneren Zusammenhang". For the nouns (about 100), however, this was not the case, and he believes, therefore, that their co-appearance in the category of greatest frequency was a matter of chance.

It should be first observed that instead of taking the results of the count by Kaeding simply as facts, Glinz acts on the implicit hypothesis according to which words of a given grammatical category, and within a specified range of frequency, must be connected by their meaning! Finding that this was not the case for nouns—and he forgot, of course, that it is only in the case of nouns, and perhaps verbs, that one could speak of a "field" of content—, he blames the statistical method for it. But rightly understood, the situation is simply this, that the results of an objective word count are not in agreement with Glinz's subjective impression, according to which they "ought" to belong to one field of meaning. However, no statistician has ever put forward such a claim and it is evident that when Glinz calls the statistical facts "fictitious", it is merely because they contradict his own prejudices.

The results of statistical linguistics must not be blamed if they do not agree with pre-conceived ideas in the matter. It is their virtue to show up wrong hypotheses.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

II

THE DISTRIBUTION LAW OF LINGUISTIC VARIABLES

2.1 LOGNORMALITY OF THE WORD-COUNT DISTRIBUTION

Although there is no absolute necessity for establishing the mathematical law which governs the word count distribution, yet it would be of considerable advantage if we could do so. To put it differently, although there is no need to superimpose on the empirical results of word counts the methods of mathematical statistics, yet this will often be found very helpful.

One great advantage of using such methods is that they may save a certain amount of arithmetical labour. As pointed out in Sect. 17.3 of Vol. 1, the use of mathematical formulae relieves us of the necessity of mere enumeration which is the only method of descriptive statistics. More especially, the use of certain limiting functions of the calculus of probability, which we call Laws of Chance, does away with the necessity of having to calculate in each case afresh the numbers and relative frequencies of combinations which we require for the evaluation of probabilities.

Several suggestions as to the distribution law governing word counts have been put forward, among which Yule's of the law of rare events, the Poisson Law, as applying to linguistic distributions, is the best founded. But apart from the fact that the pure chance mechanism of the Poisson Law does not really explain what happens in speech or writing, it is not very useful from the labour-saving point of view. The unsuitability of the Zipf hypothesis for theoretical purposes can be inferred from the preceding sections.

(1) We propose to show that word count distribution may be fitted satisfactorily by the lognormal law, which not only brings them into line with the distribution of word length and sentence length (Sects. 2.4, 9.1), but is of great advantage from the practical angle.

One of the most important of the laws of chance mentioned above is the Normal or Gaussian law, according to which the probability density $p(x)$ is given by

$$p(x) = \frac{1}{\beta\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\alpha}{\beta}\right)^2}$$

$$= \frac{1}{\beta\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\alpha}{\beta}\right)^2\right\} \qquad (7)$$

($e = 2.71828 \ldots .$
$\pi = 3.14159 \ldots .$)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

where α and β are the arithmetic mean and standard deviation, respectively.

The graph of the function shows a bell-shaped symmetrical curve with the greatest probability density around $x = \alpha$, that is around the arithmetic mean of x. Two values of x which are equally distant from the mean on the positive and negative side have the same probability density $p(x)$ which value diminishes rapidly with distance from the mean (Vol. 1, Sect. 16.4).

The cumulative distribution of the normal law (that is the sum of individuals measuring less than a stated value of the variable plotted against that value) is an *s*-shaped (sigmoid) curve which can be transformed into a straight line by plotting it on a grid whose horizontal ordinate is equidistant and whose vertical ordinate is spaced according to the error function (probability paper). On such a grid the arithmetic mean is obtained by projecting on to the base line the crossing point of the straight line graph for the distribution with the horizontal at $P(x) = .50$, and the standard deviation as the difference between that point of the base line and the projection of the crossing point of the straight line graph with the horizontal erected at either $P(x) = .84$ or $P(x) = .16$.

Apart from the distribution of individuals, we may be interested in that of the product of values of x by the corresponding frequency f_x, and, in general, of powers of x by the frequencies f_x, that is in the distributions of xf_x, x^2f_x . . . . and so on. These are called the *moment distributions of x.*

Symmetrical distributions are not the rule in nature; more often distributions are skewed, that is humped to the right or to the left. A certain class of these skew distributions can, however, be transformed into symmetrical ones by using a logarithmic transformation of the variable, with all the advantages normality of distribution entails in theoretical and practical work. Such variables are called *lognormal.* The transformed distribution, that is the logarithmic distribution, is then a normal distribution with all the properties of the normal law.

According to the lognormal law, it is not *differences* of equal amount in excess or defect from a mean value which are equally likely, but *ratios* of equal amount in excess or defect from a mean value. Let $x_1, x_2, x_3 \ldots . x_n$ be a set of values of a chance variable, and $y_1, y_2, y_3 \ldots . y_n$ be the transformation $y = \log x$. Then if

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \tag{8}$$

where μ is the arithmetic mean of y and σ its standard deviation, we say that the variable x is lognormally distributed.

The cumulative distribution of a lognormal variable can again be transformed into a straight line by plotting the cumulative percentage up to a certain value of the variable against that value of the variable on logarithmic probability paper, where the horizontal ordinate is logarithmically spaced, and the vertical ordinate according to the error function.

Conversely, if we have an empirical distribution whose cumulative probabilities

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

plot as a straight line on a log probability grid, the conclusion is justified that the variable is lognormal. We leave for the moment the question of more refined tests of lognormality.

TABLE 5

New Testament Word Count (less the last 11 words with frequency between 1716 and 19,734)

Occurrence frequency of particular words x	*Vocabulary*		*Occurrences*	
	Cumulative Frequencies $\sum_{1}^{k} f_x$	Cumulative Percentages $100 \sum_{1}^{k} f_x \Big/ \sum_{1}^{\max} f_x$	Cumulative Frequencies $\sum_{1}^{k} xf_x$	Cumulative Percentages $100 \sum_{1}^{k} xf_x \Big/ \sum_{1}^{\max} xf_x$
1	1934	35.6	1934	2.1
2	2776	51.1	3618	4
3	3246	59.7	5028	5.6
4	3568	65.7	6316	7
5	3789	69.8	7421	8.3
6	3962	73	8459	9.4
7	4085	75.3	9320	10.4
8	4195	77.3	10200	11.3
9	4302	79.2	11163	12.4
10	4381	80.7	11953	13.3
11	4464	82.2	12866	14.3
12	4528	83.4	13634	15.1
13	4570	84.2	14180	15.7
14	4622	85.1	14908	16.5
15	4664	85.9	15538	17.2
16	4703	86.6	16162	17.9
17	4736	87.2	16723	18.6
18	4772	88	17371	19.3
19	4791	88.3	17732	19.7
20	4809	88.6	18092	20.1
21	4828		18491	
22	4849		18942	
23	4869		19413	
24	4883		19749	
25	4900	90.3	20174	22.4
26	4917		21616	
27	4931		20994	
28	4945		21386	
29	4962		21879	
30	4971	91.6	22149	24.6
up to 40	5055	93.1	25081	28
up to 50	5124	94.4	28180	31.4
up to 1619	5425	100	89883	100

(2) Applying this idea to the vocabulary distribution of the whole New Testament, we obtain from the distribution as given in Table 2 (col. 2) the cumulative percentage

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

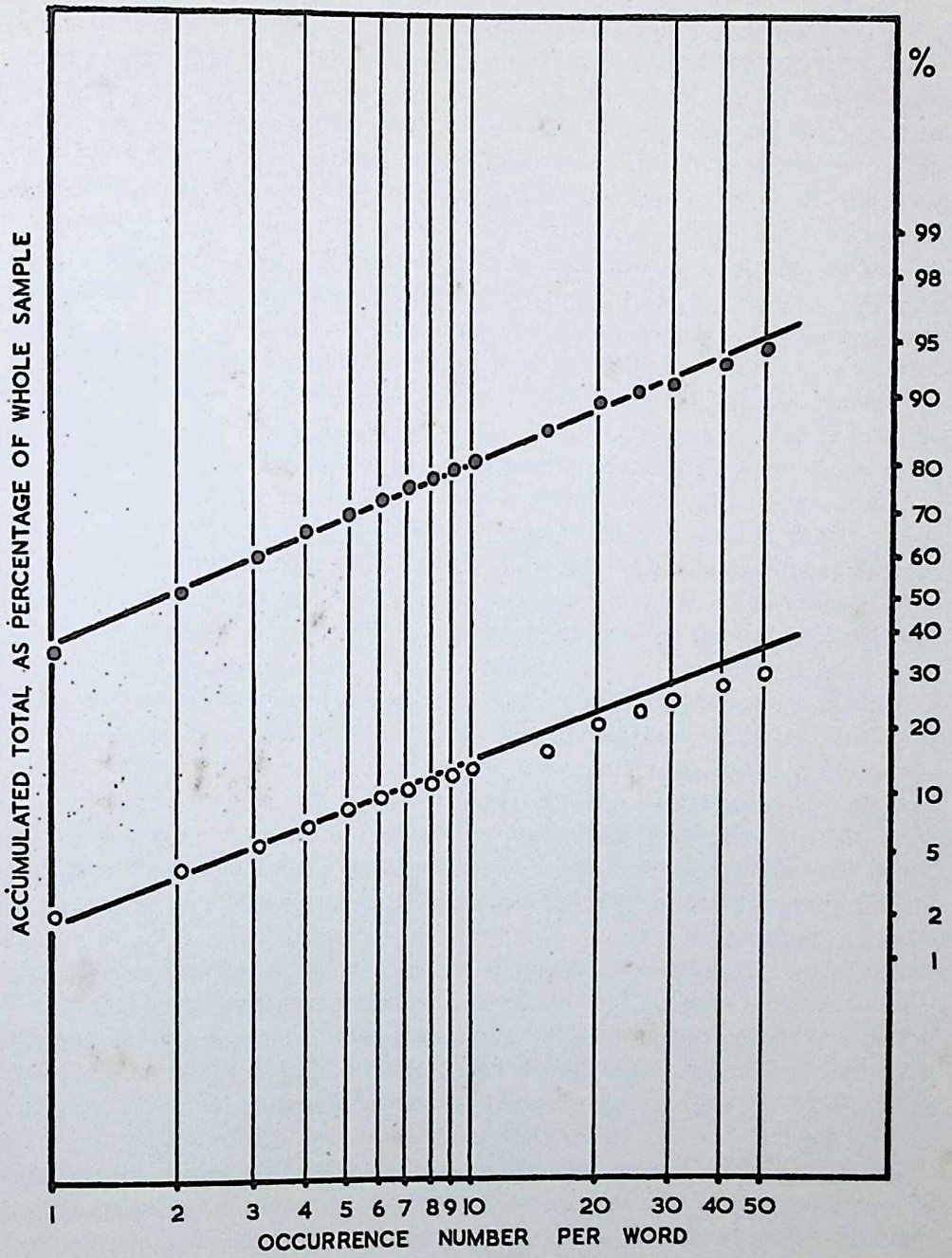


FIG. 3. New Testament word count from table 3 (•—• vocabulary, O—O occurrences).

distribution shown in Table 5 (col. 3) which, plotted on log probability paper, gives a very satisfactory straight line graph (Fig. 3).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

A certain adjustment of the occurrence data appeared indicated. The last 11 words in the order of ascending frequency are all grammar words with frequencies ranging from 1,716 to 19,734 (for the articles). Altogether these 11 words account for 47,445 occurrences. Since the type-token relation for these vocabulary items (11 vocabulary items to 47,445 occurrences) is quite out of keeping with the type-token relation for the vocabulary as a whole, and its inclusion would therefore have the effect of disturbing the general relation and deprive it, to some extent, of its peculiarity or individuality, it was decided to use as the denominator of the cumulative percentage of occurrences 137,328 — 47,445 = 89,883.[1] This type of adjustment is generally recommended if we wish to make use of the lognormality of word counts.

If the series is a short one ($n < 30$, say), μ and σ ought to be determined numerically. For a long series with a range from 1 to over 1700 occurrences, as we have it in the adjusted series of occurrences, the calculation of μ and σ would be so laborious as to offset largely the benefits to be derived from the established lognormality of word counts. In such cases it is preferable to use the graphical method explained above. Another reason in favour of the graphical determination of σ is that, whereas in the beginning the series proceeds in steps of unity, the later part may show gaps of 1,000 units or more, which makes a consistent grouping of the observations very difficult; any change of interval size introduced along the range must needs affect the correctness of the statistics.

Absolute precision is not obtainable in word counts for a variety of reasons, nor is it really necessary or even useful for the purpose of establishing style or language characteristics. These are matters where we can never dispense with at least a modicum of subjective judgement when all the statistical evidence has been marshalled.

(3) The hypothesis of lognormality receives support from the nature of the distribution of occurrences according to the frequency of vocabulary items, and still more from the relation in which the two distributions, of words and of occurrences, stand to one another.

One of the most interesting properties of the lognormal distribution is that the j'th moment distribution of a lognormal distribution with mean μ and standard deviation σ is also a lognormal distribution with mean $\mu_j = \mu + j\sigma^2$ and standard deviation σ, respectively. It follows that the graphs of different moment distributions should be parallel lines on log probability paper with a distance of $j\sigma^2$ between their means.[2]

The theoretical difference between the moment variables as given above is in terms of natural logarithms, that is logarithms to the base e. If logarithms to the base 10 are used, the following transformation must be observed. Since the mean and standard deviation of the log distribution are the logarithms of the geometric mean G and the geometric standard deviation, σ_G, respectively, viz.

$$\mu = \ln G, \quad \sigma = \ln \sigma_G,$$

[1] Correspondingly we have for the vocabulary 5436−11=5425 words.

[2] For the proof see Appendix A.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 6

	Logarithmic mean (base 10)	Difference	Logarithmic S.D.[3] (base 10)	Average S.D. ($\bar{s}$)	2.3026 $\bar{s}^2$
New Testament					
Vocabulary	0.2553	1.7447	0.8587	0.8800	1.783
Occurrences	2.0000		0.9031		
Essay on Bacon					
Vocabulary	0.1461	0.7290	0.5863	0.5910	0.80
Occurrences	0.8751		0.5960		

the relation

$$\mu_j = \mu + j\sigma^2 \tag{9}$$

can be written as

$$\ln G_j = \ln G + j \ln^2 \sigma_G$$

Transforming into logarithms to the base 10, we have

$$\log_{10} G_j = \log_{10} G + 2.3026 j \log_{10}^2 \sigma_G$$

or

$$\mu_j = \mu + 2.3026\, j\sigma^2 \tag{10}$$

The distribution of occurrences as the *first* moment distribution of the distribution of vocabulary should, therefore, plot as a straight line on logarithmic probability paper and, since $j = 1$, be σ^2 distant from the curve for the vocabulary distribution. As can be seen from Fig. 3 this is actually the case.

The quantities required for the comparison are listed in Table 6.

The last column shows $2.3026\sigma^2$ with the mean value, s, from the preceding column substituted for σ. These figures are in good agreement with the observed figures given in the third column. We may, therefore, take it as a reasonable hypothesis supported by statistical evidence that the distributions of vocabulary and of occurrence against frequency of occurrence of particular words satisfy the criteria for lognormality, by themselves and in their relation to one another, since the graphs of the cumulative distributions, columns 3 and 5 of Table 5, are sensibly straight lines and the difference between the means of the vocabulary and the occurrence distribution is approximately equal to $2.3026\sigma^2$ (Fig. 3). We may, therefore, take the lognormality of word count distributions as established in this particular case, but we shall have plenty of opportunity for satisfying ourselves again and again on this point.

[3] Standard deviation

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

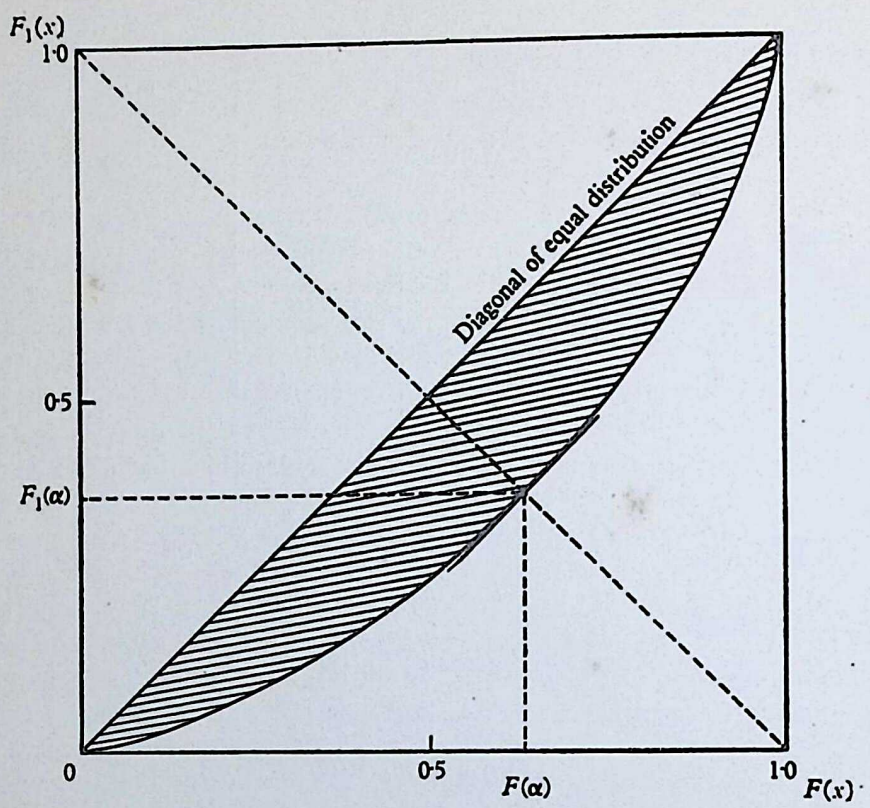


FIG. 4. The Lorenz diagram for a two-parameter distribution. This curve with index $L=0.383$ is derived from $\Lambda(\mu, 0.5)$ (Aitchison and Brown).

As just another illustration of the lognormal property of word counts, Fig. 5 shows the graphs of the frequency distribution of *nouns only* in a sample of about 8000 words from Macaulay's *Essay on Bacon* (Table 8a) for both vocabulary and occurrences (full lines, large dots). The characteristic moment distribution property of the latter in relation to the former is confirmed by the quantities listed in Table 6.

2.2 THE MEASURE OF CONCENTRATION FOR WORD COUNTS

It is common practice in word count analysis to relate vocabulary and occurrence by way of stating the number of different words which account for specified percentages of the total number of occurrences in the sample, e.g. for the New Testament we can, for instance, say that the 10 most frequent words account for about 45,000 occurrences and 170 vocabulary items with more than 100 occurrences each for about 100,000 occurrences.

If it is desirable to express the type-token relation for a given sample or text more succinctly, we use the Characteristic K or v_m (Sect. 5.1), which is independent of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

sample size. This makes no assumption whatever about the mathematical nature of the distribution function, but it requires quite a bit of computational work.

As observed above, it is one of the great advantages of establishing the distribution law governing empirically obtained data that it might save computational work. Having found that the lognormal distribution provides a satisfactory description of word counts, we shall now show how a general characteristic (not K, though) of the type-token relation can be obtained with a minimum of arithmetic.

(1) The correlation of the number of words (vocabulary items) with that of their occurrence in the text could be systematically presented by the method of concentration leading to the so-called Lorenz diagram.

In the Lorenz diagram (Fig. 4), if applied to word counts, the proportion of *words having less than x occurrences* per word is measured along the horizontal scale and the *proportion of occurrences* accounted for by those words along the vertical scale. In other words: the cumulative percentage distribution of the amount of the variable concerned is plotted as ordinate against the cumulative percentage distribution of the individuals possessing that amount. The points plotted for the various values of x describe a curve below the diagonal inclined under 45° sloping upwards to the right from the origin. In statistical terms, the curve describes the relation between the distribution function $F(x)$ and the first moment distribution $F_1(x)$ defined by

$$F_1(x) = \int_0^x x dF(x) \Big/ \int_0^\infty x dF(x) \tag{11}$$

The measure of concentration which is naturally suggested by the Lorenz diagram is the ratio of the shaded area between the Lorenz curve and the diagonal to the area of the triangle under the diagonal. That measure varies from zero, when the proportion of occurrences is numerically, for every value of x, the same as the proportion of vocabulary items to which they belong, in which case the curve is identical with the diagonal of *equidistribution*, to unity when the curve extends and fills up completely the triangle below the diagonal. The formal definition of the measure is

$$L = 1 - 2 \int_0^\infty F_1(x)\, dF(x) \tag{12}$$

Substituting in equation (11) the expected form for F_1 we obtain under the lognormal hypothesis (Proof given in Appendix A):

$$\begin{aligned} L &= 1 - 2 \int_0^\infty \Lambda (x \mid \mu + \sigma^2, \sigma^2)\, d\Lambda (x \mid \mu, \sigma^2) \\ &= 1 - 2\Lambda (1 \mid \sigma^2, 2\sigma^2) \\ &= 1 - 2N \left(\frac{-\sigma}{\sqrt{2}} \,\middle|\, 0,1 \right) \\ &= 2N \left(\frac{\sigma}{\sqrt{2}} \,\middle|\, 0,1 \right) - 1 \end{aligned} \tag{13}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

This means that for the lognormal distribution the *Lorenz factor depends only upon the value of the logarithmic standard deviation*, σ and can be read off immediately from a numerical table giving values of L for specified values of σ (App. A1). This result is of great importance for the analysis of word count distributions. It is a direct consequence of the property of lognormal distributions — also very useful for word count analysis — according to which the moment distributions are again lognormal with parameters $\mu' = \mu + j\sigma^2$ and σ (Sect. 2.4).

(2) Wishing to apply this result to the New Testament word counts, with a minimum of arithmetical labour, we first find the standard deviation from the lognormal probability graph by projecting the crossing point of the line for the word count with (a) the ordinate at 50% and (b) at 16% or 84% on to the base line. We find these points as 1.8 and 13 respectively. The difference of the logarithms of these quantities is 0.8587 and represents the logarithmic standard deviation σ. In order to use Table A1 which gives the natural logarithm, we must express σ also in terms of the natural logarithm. Multiplying it by 2.3026, we obtain it as 1.98. From A1 we read the Lorenz measure for this value of σ as 0.8415. Similarly, for the *Essay on Bacon* where σ is approximately 0.6, and, in terms of $\log_{10}$, 1.38, for which we read the Lorenz measure as 0.6778. This shows the vocabulary of the *Essay on Bacon* as being less highly concentrated than that of the New Testament.[4]

L is algebraically not idential with K or v_m, but serves the same purpose. Provided our hypothesis about the lognormality of word count distributions is, in general, correct—and we shall find that there can be little doubt about this—the arithmetical labour for calculating K as a measure of concentration can be saved by obtaining the Lorenz factor L in the manner described above.

2.3 ESTIMATING THE WORD COUNT DISTRIBUTION ON THE BASIS OF THE LOGNORMAL HYPOTHESIS

The following is an instance where the assumption of lognormality of word counts saves us the trouble of having to make a new word count.

As material we use Yule's count of nouns in 4 samples from Macaulay's *Essay on Bacon*. Yule collected by stratified sampling, 4 samples of approximately 2000 occurrences (nouns only) each from non-overlapping parts of the text. Next he combined any two of the samples into samples of approximately 4000 occurrences, then he combined any three of them into samples of approximately 6000 occurrences and ultimately he built all the observations into one sample of approximately 8000 occurrences. Table 8a gives, in columns 2-5, the word count distribution for the original samples of 2000 nouns each; of the combined samples of 4000 and 6000 occurrences table 8a gives specimens in columns 6 and 7 respectively; the last column contains the distribution of vocabulary in the total sample of about 8000 nouns.

[4] This is partly due to the former sample being one of nouns only.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 8a *Word counts from Macaulay's essay on Bacon.*

Frequency of occurrence of particular words x_k	Number of different words (Vocabulary) f_x						
	a	b	c	d	cd	bcd	abcd (Total)
1	535	580	589	591	778	906	990
2	169	174	158	167	253	311	367
3	69	68	68	73	124	162	173
4	42	29	32	47	66	87	112
5	24	24	21	25	49	58	72
6	18	15	16	24	34	45	47
7	14	9	15	11	31	46	41
8	4	8	12	8	11	22	31
9	7	8	7	2	11	19	34
10	5	5	3	4	15	14	17
11	3	3	5		12	14	24
12	4	4	1	3	7	13	19
13	1	2	1	3	6	11	10
14	2		2	1	6	10	10
15		2			4	4	13
16	1			1	3	3	3
17	2				2	7	10
18		2	3	1	3	4	7
19			1		1	4	6
20	1		1		4	3	5
21				1	3	6	1
22			1	1	2	1	4
23		1			1	4	7
24						2	2
25						5	1
26	1					2	5
27					1	1	3
28						2	4
29		1				2	1
30					1	1	3
31						1	2
32						1	1
33						1	1
34							1
35							1
36					1	1	1
37							1
38							2
39					1		4
40					2		1
41						1	1
45						1	2
48							1
53			1				
54						1	

Vidya Niwas Mishra
University of Gorakhpur,
GORAKHPUR.

डा० विद्या निवास मिश्र
प्राध्यापक, भाषा विज्ञान
निजी पुस्तकालय
पुस्तक संख्या.....

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Table 8a, continued

Frequency of occurrence of particular words x_k	Number of different words (Vocabulary) f_x						
	a	b	c	d	cd	bcd	abcd (Total)
55						1	
58							1
62						1	
63	1						
64		1					
65							1
69						1	
75				1			1
76							1
81							1
89							1
128					1		
192						1	
255							1
Sum	903	936	937	964	1432	1780	2048

Let us suppose we had counted only one sample of about 2000 words from Macaulay's essay on *Bacon*. Plotting the number of words against the frequency of occurrence on a lognormal grid, we get a sensibly straight line (line b, Fig. 5). Let us further suppose we wanted now the vocabulary distribution in a sample of twice that size, or three times that size, etc. As we know, the vocabulary number will increase with sample size, but by no means proportionately to it: not only will each frequency class of words have its own rate of increase, but all these rates will diminish as the sample is further increased. Renewed counting thus seems the only way for obtaining the new word count distribution, were it not for the established lognormality of such distributions.

We first remember that the word count in a sample whose size is a multiple of 2000 is just another word count distribution, and, as such, will again be a straight line on the log probability grid. But where is that line to be, and what direction is it to have?

Considering now that the concentration of the type-token relation epitomised by Lorenz is independent of the size of sample and dependent only on σ (formula 13), it follows that all distributions having the same concentration in terms of the Lorenz factor must have the same σ, and since every sample from the total text is, by supposition, characterised by the same type-token relation, it follows that all samples will have the same σ.

We then remember that probability graphs of homoscedastic normal distributions,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

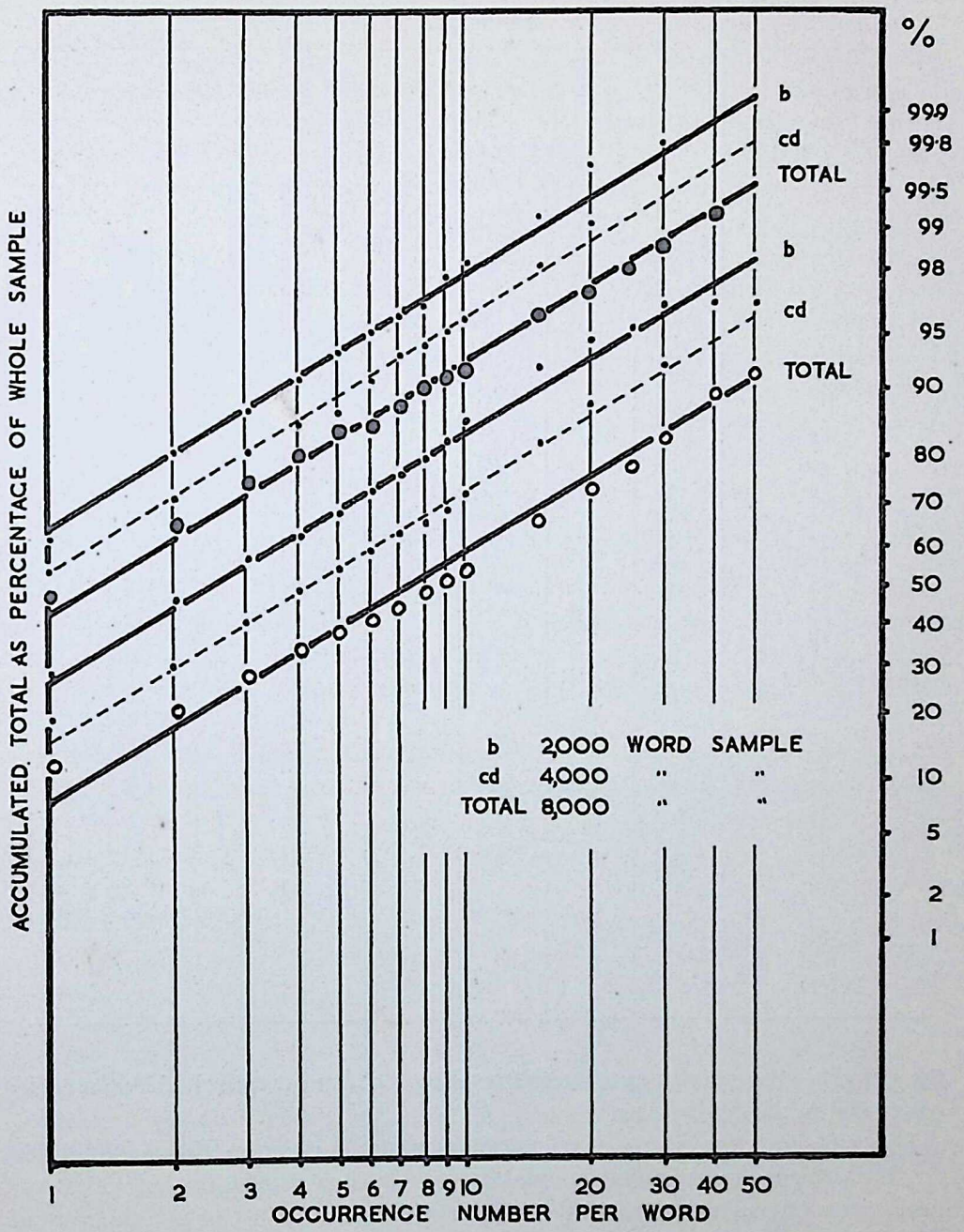


FIG. 5. Macaulay ["Bacon" Essay] word count from table 8 (•—• vocabulary, O—O occurrences).

that is distributions with the same σ, are parallel lines on a probability grid, and for lognormal distributions on a log probability grid, and this determines the direction of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 8b

Frequency of occurrence of particular words	*Vocabulary* Cumulative frequency: $100 \sum_{1}^{k} f_x \Big/ \sum_{1}^{\max} f_x$				*Occurrences* Cumulative frequency: $100 \sum_{1}^{k} xf_x \Big/ \sum_{1}^{\max} xf_x$			
x_x	*b*	*cd*	*bcd*	Total	*b*	*cd*	*bcd*	Total
1	62.0	54.2	50.2	48	29.0	19.2	15.4	12.3
2	80.5	72	68.4	66	46.5	31.6	26	21.4
3	87.8	80.6	77.4	75	56.8	40.8	34.1	28
4	91	85.3	85.6	80	62.5	47.3	40	33.5
5	93.4	87	88.1	84	68.5	53.4	45	38
6	95	91	90.7	85	73	58.3	49.5	41.4
7	96	93.2	92	88	76.2	63.3	55	45
8	96.3	94	93	90	79.4	66	58	48
9	97.8	95	93.8	91	83	68.4	60.8	52
10	98.2	95.8	94.6	92	85.5	72	63.2	54
11	98.6	96.6	95.3	93	87.2	75.3	65.8	57.3
12	99	97	96	94	89.3	77.3	68.4	60
13	99.2	97.5	96.5	94.3	90.9	79.3	70.8	62
14	99.2	98	96.7	95	90.9	81.3	71.5	63.5
15	99.2	98.2	97	96	92.4	82.8	72.5	66
16	99.2	98.5	97.3	96	92.4	84	73.3	66.4
17	99.2	98.6	97.5	96.3	92.4	84.9	75.3	68.4
18	99.7	98.7	97.7	97	94.2	85.8	76.5	70
19	99.7	98.7	97.8	97	94.2	86.2	77.4	71.5
20	99.7	99.1	98	97.2	94.2	88.1	78.7	72.8
up to 25	99.7	99.5	99	98	95.3	91.5	85.6	77
„ 30	99.8	99.6	99.3	98.7	96.7	92.3	89.2	82.5
„ 35	99.8	99.7	99.5	99	96.7	93.1	90.8	85
„ 40	99.8	99.9	99.5	99.2	96.7	96.8	92.1	89.3
„ 50	99.8	99.9	99.7	99.2	96.7	96.8	92.8	91.5
(64)	100				100			
„ 70		99.9	99.9	99.8		96.8	96.8	93.7
(128)		100				100		
(192)			100				100	
(255)				100				100

the straight lines representing samples from Bacon of size other than 2000: they will be parallel to the original line.

Figure 5 shows the scatter of points representing the samples of 4000 and 8000 words as being very satisfactorily fitted by straight lines parallel to that which represents the original sample of 2000 nouns, as predicted.

This suggests that from the graph of a sample of a given size we can easily get estimates of the distribution of vocabulary in samples of greater, or smaller, size. In order to obtain an approximate estimate of that sort we make use of the property of word count distributions, according to which the new lines will be parallel to the original one on a log probability grid. This determines their *direction*. In order,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

however, to construct such a line, we must know in addition, at least one point which would give us its *position*. The most reliable point for this purpose is the starting point of the line on the graph given by the relative frequency of the hapax legomena, i.e. words occurring only once in the sample. For the 4000 word sample (*cd*) that relative frequency is 54.2% for the vocabulary, and 19.2% for the occurrence distribution; for the 8000 word sample it is 48% for the vocabulary and 12.3% for the occurrence distribution. Through these points we draw lines parallel to the original graph. These represent the cumulative frequency distributions in samples twice, three times, etc. the size of the original, and working backwards from the graph we obtain an estimate of the probability density distribution of our variable. Lacking precise knowledge of how the number of hapax legomena increases with sample size, we must use a rule-of-thumb measure derived from the empirical knowledge that the ratio of once-words to total occurrences decreases at a slower and slower rate as the sample size increases.

2.4 SENTENCE LENGTH AS A STYLE CHARACTERISTIC

2.4.1. Sentence length as a statistical (lognormal) variable

The length of sentences in terms of the number of words is, to a considerable extent, a matter of individual choice, and thus of style, rather than language structure. Accordingly, we find characteristic differences between authors in this respect, and a far reaching conformity in text samples from the same author. Qualifications of these statements will be found below.

Sherman (*Analytics of Literature*, Boston, 1893) gives the average sentence length in terms of the number of words in samples of 500 from different authors as follows:

Fabyan	63.02
Spenser	49.82
Hooker	41.40
Macaulay (essays)	22.45
Channing	25.73
Emerson	20.58

He then compares 22 samples of 100 sentences each by de Quincey and obtains an average of 33.25, and fairly good conformity between the 22 samples. He further compares Macaulay's *History of England* with his *Essays*, and he obtains for 4 samples of 10,000 sentences each from the former, the following average sentence lengths: 23.33, 23.18, 22.32, 23.73. The total number of periods in the *History of England*, he gives as 41,579; in the remaining 41,579 — 40,000 = 1,579 sentences, there are 38,696 words, which gives for the remainder 24.5 words per sentence. The total average results as 23.41.

The distribution of sentence length according to the number of words is, as a rule, a skew distribution (skewed to the left) and was found by Williams (1940) to assume a more or less symmetrical form if a geometrical or logarithmic scale was used for the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

number of words per sentence. These distributions were shown to be of the log-normal type.

2.4.2 *Sentence length and authorship*

However, as in the case of using other style characteristics, we must distinguish between the use of sentence length for characterising style, and as a criterion of authorship. Yule (1939) who discussed sentence length as a statistical characteristic of style, came to the conclusion that differences between samples from texts by the same author may greatly exceed the limits of simple sampling, without for practical purposes being of any real, i.e. literary, significance. To put it succinctly, *philological or literary identity of authorship is not always reflected by statistical identity of sentence length*, within the limits of sampling fluctuations, and, conversely, *literary diversity is not always reflected by statistical diversity* in terms of significant differences in sentence length.

In general, sentence length decreases with time (see below for numerical evidence) and insofar is a language-conditioned feature, and unfit to serve as a criterion of authorship. It follows that the comparison of sentence length with a view to characterising individual style should be confined to contemporaneous writers.

Moreover, it is evident that there may also be short-term changes of style within the lifetime of a single author, and changes of style according to the subject he is writing about.

The problem of the stability of the sentence length distribution in different works of a given writer was made the subject of an investigation by Buch (1949). He used two works of the Danish author, Johannes V. Jensen, published with an interval of about 20 years, viz. *Den lange Rejse*, which appeared during the years 1908-22 and *Myter*, vol. 8 (1940). From each of the two works were chosen arbitrarily 600 sentences, and the number of words in each of these was counted. How the selection of sentences has been made is of no importance in this connection.

He found both distributions to be in good conformity with the *hypothesis of lognormality*. The following table gives the statistics of the arithmetic and logarithmic distribution, where the Greek letters stand for the arithmetic mean and the standard deviation of the lognormal distribution.

Applying the Kolmogoroff-Smirnoff test for homogeneity of the two distributions, he obtained the result that the differences were not compatible with the assumption of their being both random samples of one statistical population. The obvious interpretation is that the author's style has undergone, in the course of time, so great alterations that the two works could not be regarded as belonging to the same category as far as sentence length distribution is concerned.

These results contradict Wake's (1957) conclusions about sentence length as an objective criterion of authorship. The hypothesis of sameness and difference of authorship as being reliably reflected by non-significant and significant differences respectively of sentence length is not tenable. This is shown by his own data on

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 9

	Den lange Rejse	*Myter*
Number of sentences	600	600
Number of words	21821	17802
Arithmetic mean of words per sentence	36.37	29.67
Logarithmic mean μ	1.417	1.358
Logarithmic standard deviation σ	0.383	0.342

sentence length in works by Xenophon, Aristotle and Plato. Regarding the question of stability of sentence length in different works by the same author, he admits that his samples of works by Plato show greater differences in the mean sentence length than could be accounted for by random sampling. Conversely, the grand mean sentence length for samples from a number of works by Xenophon results as 18.9 with a standard error of 0.4, and that of samples from works by Aristotle as 18.4 with a standard error of 0.8, which, if his conclusion were correct, would mean that the works by Xenophon and Aristotle have the same author. Even if Yule's conclusion about the differences in mean sentence length between different samples from the same text greatly exceeding the limits of simple sampling may have been incorrect in a single instance, as Wake states, in general Yule's statement is a better assessment of the stability of sentence length than Wake's. Only in connection with other independent evidence can sentence length be useful for establishing sameness or difference of style, and with it throw light upon questions of disputed authorship.

2.4.3 Language-conditioned elements of sentence length

The change of sentence length with time as a long range trend from longer to shorter sentences was investigated by Sherman (1893). Sherman compared older and newer English texts with regard to the number of sub-clauses or the extent of predication, and found it to be much greater in older English. Here then is a hint of the reason for the decrease in numerical sentence length in English. According to Sherman, it is not so much the splitting up of the long sentences occurring in older English texts, but rather the abandoning of predication in favour of more syncopated construction, which accounts for the reduction of sentence length.

The following are the results of a count of predications per sentence in 500 sentences from each writer, and of the percentage of simple sentences.

It is quite evident that predication decreases with time, whereas the percentage of simple sentences increases. This bears out the assumption that the decrease in sentence length was due mostly to "internal" changes, viz. the abandoning of clauses or predication in favour of participle construction or other substitutions for the finite verb.

Predication conditions the imagination and in prose imposes upon the faculty of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 10

Author	Average number of predications	Percentage simple sentences
Chaucer	5.24	4
Spenser	4.68	11
Sidney	3.98	10
Hooker	4.12	12
Bacon	3.12	19
Bunyan	3.94	10
Addison	3.67	12
Bolingbroke	3.72	13
de Quincey	3.69	14
Macaulay	2.15	39
Channing	2.56	31
Emerson	2.36	37
Grant	2.36	31
Everett	2.39	32
Bartol	2.00	45

imagination the obligation of realising with exactness the facts of the relations affirmed. On the other hand, that style which leaves most to fancy in respect of the manner in which facts or relations may be apprehended will be insofar the easiest to read. *The prevailing fault of the synthetic style is the affirmation of the obvious* (Sherman).

In terms of "information", this means that the older writer liked to cross his *t*'s and dot his *i*'s when expressing an idea, whereas the modern author leaves more to fancy and imagination in respect to the manner in which facts or relations may be apprehended. He writes, as we say, a less heavy, in the sense of "ponderous", style, and though it requires closer attention and greater speed in thinking on the part of the reader to follow him, yet he is easier to read.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

III

THE STATISTICAL VERSION OF DISCOURSE ANALYSIS
THE DISTRIBUTION OF VOCABULARY OVER SEVERAL SAMPLES

The technique of discourse analysis as developed by Z. Harris (1952) is, as a rule, applied to individual words and sentences. It has two aspects: (1) the distribution of the same lexical item in various contextual situations within the text; and (2) the distribution of different lexical items in identical or similar contextual situations within the text.

The following considerations lead naturally to the statistical version of discourse analysis. Recording the occurrence of the same lexical item in different contexts of a given text is of the nature of establishing a *concordance*, and this is also the term which Longacre (1958) uses for that aspect of discourse analysis. True, a concordance will record *all* occurrences of a vocabulary item, regardless of whether it is used in the same sense or not. However, if a word occurs in different parts of the text, the probability that it will be used with different meanings increases with the number of such occurrences, and the difference between concordance in the conventional sense and in the sense in which the term is used in discourse analysis is, thus, only apparent. In any case, since it is, at the most, a difference in degree and hardly definable, we may regard a statistical dictionary of the kind described in Sect. 1.3 which gives for each vocabulary item in the Greek text of the New Testament the number of occurrences in the various books of the text as an example of the *integral* form of the first aspect of discourse analysis.

The second aspect of discourse analysis, the distribution of different lexical items in associated (through the part they have in common) contexts, is called by Longacre the equivalent-chain relation. A better term, which we shall use, would be "*concatenation*" of contexts, since as we proceed to chains of higher order, the aspect is that of the distribution of vocabulary items according to whether they are peculiar to a given context, or common to 2,3 ... *n* contexts.

As in the case of "concordance", the analysis in its original form is here again concerned with individual sentences selected from the text. However, in the actual application of discourse analysis, use is made of grammatical transformations in order to secure more easily comparable contexts. Further, two portions of a text are considered to be identical for the purpose of analysis if they contain the same lexical item as head. Finally, a substantive expression and its pronominal substitute are considered equivalent. The need for such simplifications and neglect of minor

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

differences between contexts leads naturally, if done on a larger scale, to the type of generalisation implied in statements of a statistical nature. Integration over the whole text produces a complete picture of context concatenation, or vocabulary connectivity, in the text.

The advantages of proceeding from the discourse analysis of individual passages to the integrated form are these:

1. The more or less anecdotal, and partly arbitrary, stage of discourse analysis as practised gives way to the statistical theory of the distribution of vocabulary over several samples, which provides a complete picture of vocabulary connectivity in a perfectly reproducible manner.

2. The mathematical structure of vocabulary partitioning among different contexts of text samples provides us with a measure, or standard, of the distribution of vocabulary in both aspects of discourse analysis, against which we can judge whether in a given case the distribution deviates from what could happen by mere chance and, if so, how great such deviation is.

In this chapter, the statistical theory of vocabulary partitioning is set forth. Chapts. VI and XVIII show the application to problems of disputed authorship.

3.1 MATERIAL AND NOTATION

In order to arrive at an objective description of the phenomenon which forms the title of this chapter, we shall use word counts of nouns only by Yule (1944) made for quite a different purpose, but very suitable for our problem. Our material will consist of three series of data: 4 spatially different samples from the same work, Macaulay's essay on *Bacon*; samples from 4 different essays by Macaulay—on Milton, John Hampden, Frederick the Great, and the total count from the essay on Bacon; and samples from 4 works by Bunyan—*The Pilgrim's Progress*, Part I, *The Pilgrim's Progress*, Part II, *The Life and Death of Mr. Badman*, *The Holy War*. Since the first group of 4 samples represents one particular text, the second group 4 different texts but of the same literary genre, and the third 4 different works by the same author, there is sufficient variation in the relation of smaller to larger unit to warrant reasonable generality for our conclusions.

We shall use the following notation. The capital A (positive symbol) denotes "noun occurring in sample A"; the corresponding Greek letter α (negative symbol) "not occurring in sample A", and so on. A class symbol like $ABCD$ denotes "occurring in all 4 samples, A, B, C and D"; $ABC\delta$ "occurring in A, B and C, but not in D"; $A\beta C\delta$ "occurring in A and C, but not in B and D", and so forth. There are 4 letters in each class symbol of Table 11, and each letter may be either a capital or a Greek letter, so there must be in all 2^4, or 16 classes.

A class frequency is denoted by putting the class symbol in brackets, e.g.

$$(\alpha\beta\gamma\delta) = 0$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

According to the number of letters within the brackets, we speak of first order, second order, third order and fourth order frequencies. A symbol like (*A*) is a first order frequency and means all words occurring in sample A, regardless of whether they also occur in any of the other samples; a symbol like (*BC*) denotes a second order frequency and means all the words occurring in B and C, regardless of whether they occur in *C* and *D*, etc.

Table 11 shows, in the columns headed "Observed", the distributions of the total vocabulary over the 16 word classes in the 4 samples from Macaulay's essay on *Bacon*, the 4 essays by Macaulay, and the 4 works by Bunyan. The explanation of the columns headed "Expected" will follow.

TABLE 11

Comparing the observed class frequencies of the fourth order with those expected on the basis of random partitioning (Yule, 1944)

1	2	3	4	5	6	7	8
	Class	Sampling experiment		Macaulay		Bunyan	
		Expected	Observed	Expected	Observed	Expected	Observed
1	*ABCD*	277.51	271	607.25	402	449.17	259
2	*ABCδ*	61.60	62	43.21	55	75.64	82
3	*ABγD*	61.60	58	154.17	132	75.64	92
4	*AβCD*	61.60	53	154.17	144	75.64	66
5	*αBCD*	61.60	64	154.17	140	75.64	51
6	*ABγδ*	71.58	68	70.48	79	76.02	111
7	*AβCδ*	71.58	64	70.48	100	76.02	83
8	*AβγD*	71.58	59	184.17	178	76.02	50
9	*αBCδ*	71.58	67	70.48	74	76.02	63
10	*αBγD*	71.58	66	184.17	168	76.02	65
11	*αβCD*	71.58	81	184.17	173	76.02	74
12	*Aβγδ*	273.66	268	319.12	447	259.53	262
13	*αBγδ*	273.66	280	319.12	283	295.53	297
14	*αβCδ*	273.66	275	319.12	457	259.53	352
15	*αβγD*	273.66	312	708.72	711	259.53	339
16		—	—	—	—	—	—
Total		2048.03	2048	3543.00	3543	2245.97	2246

3.2 FEATURES OF VOCABULARY STRUCTURE

Careful inspection of Table 11 reveals a remarkable similarity in the runs of frequencies in columns 4, 6, 8. In each, the classes with the highest frequencies are (1) the leading class *ABCD* of nouns common to all 4 samples, and (2) the classes *Aβγδ* etc. of nouns peculiar to a single sample, which we shall also briefly refer to as one-class words. To bring this trend out more clearly, we average the three groups of lines 2-5, 6-11, 12-15.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 12

Giving (in addition to line 1) the averages of lines 2-5, 6-11, and 12-15 of Table 11, i.e. mean frequencies of classes of words common to four, three or two of the samples or special to one only

1	2	3	4	5
Lines of Table 11	Words common to number of classes	Four samples of Essay on *Bacon*	Four essays of Macaulay	Four works of Bunyan
1	4	271.00	402.00	259.00
2-5	3	59.25	117.75	72.75
6-11	2	67.50	128.67	74.33
12-15	1	283.75	474.50	312.50

Moreover, if we now pick out from Table 11 the 8 classes of the fourth order which together make up any one class of the first order, we find that the frequencies $(ABCD)$ and $(A\beta\gamma\delta)$ taken together contribute one-half or more of the total (A), and similarly the frequencies $(ABCD)$ and $(\alpha B\gamma\delta)$ one-half or more of the total (B), and so forth. That is to say, one-half or more of all the nouns in any one sample is contributed by (1) nouns common to all 4 samples and (2) nouns subject to that particular sample. Table 3 gives the percentages of the form.

$$100\,\frac{(ABCD) + (A\beta\gamma\delta)}{(A)} \tag{14}$$

For example, from column 4, Table 11, we have

$(ABCD) = 271$

$(A\beta\gamma\delta) = 268$

$(A) = 271 + 62 + 58 + 53 + 68 + 64 + 59 + 268 = 903$ and the ratio

$$(271 + 268)/903 = 59.7\%$$

From column 6, Table 11, we get the corresponding ratio as

$$(402 + 447)/1537 = 55.2\%$$

and so on.

3.3 RANDOM PARTITIONING OF VOCABULARY

The uniformity is so striking that it looks as if we ought to be able to account for it by some purely statistical reasoning.

Let us consider a text as a linear sequence of 100,000 words, which is divided into 100 samples of 1000 words each. If a word occurs once only in the total, the probability of its appearing in one sample is .01 and, consequently, that of not appearing, .99; the probability of its appearing in one or the other of two samples is, by the addition theorem of probability, .01 + .01 = .02, and that of not appearing, .98; the corresponding probabilities in the case of three samples are .03 and .97.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 13

Giving the percentages of words common to the four samples, together with words special to each sample, on the total for each of the samples A, B, C and D of Table 11

1	2	3	4
Class	Four samples of essay on Bacon	Four essays of Macaulay	Four works of Bunyan
A	59.7 (58.1)*	55.2 (57.8)	51.8 (60.9)
B	58.9 (58.1)	51.4 (57.8)	54.5 (60.9)
C	58.3 (58.1)	55.6 (57.8)	59.3 (60.9)
D	60.5 (58.1)	54.3 (56.5)	60.0 (60.9)
Average	59.4	54.1	54.4

* The figures in brackets are the expected percentages (see end of Sect. 3.3).

In general, for r samples, the probability of the word appearing in one of k samples is k/r, and that of its not appearing in k samples

$$\frac{r-k}{r} \tag{15}$$

The most convenient set of quantities for our purpose is that of the probability of non-appearance, from which the probabilities, and expected numbers of words in the various classes, can easily be obtained.

If the word occurs twice in the whole text, the above probabilities for appearance and non-appearance in a sample apply for each occurrence of the word, and the probability of neither of the two occurrences appearing in one sample is then, by the multiplication theorem of probability, $(.99)^2 = .98$; in two samples $(.98)^2 = .96$; in three samples $(.97)^2 = .94$. If the word occurs three times in the text, the respective probabilities of its non-appearance in a sample are $(.99)^3 = .97$, $(.98)^3 = .94$, and $(.97)^3 = .91$.

Generally, for x occurrences and r samples, the probability of non-appearance in k samples is

$$\left(\frac{r-k}{k}\right)^x \tag{16}$$

If there are f_x such words (the sum of all f_x being the number of words in the total, 100,000 in our illustration), the expected number of them not appearing in any of the k samples is

$$f_x\left(\frac{r-k}{r}\right)^x \tag{17}$$

and the entire number of words not falling into any one of the k samples is the sum of such expressions for all values of x

$$S\left\{f_x\left(\frac{r-k}{r}\right)^x\right\} \tag{18}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The random distribution of vocabulary may be visualised more clearly by the following experimental demonstration. Let us suppose we write out all the nouns of the total frequency distribution of words from Macaulay's essay on *Bacon* on counters, giving a noun that occurs x times x counters, and then deal them at random into four trays labelled A, B, C and D. To ensure approximate randomness we might toss the counters into a spinning circular tray with four equal quadrants A, B, C and D. Sorting the counters in each quadrant by nouns, and listing and counting the *nouns* in each—not the number of counters—, we shall then have word classes like those of column 4, Table 11. The most convenient way is to calculate the set of frequencies for the *negative* classes, the symbols for which consist entirely of Greek letters. Supposing now our circular tray divided not into four, but into r equal sectors, the numbers *expected* on pure chance to fall into the classes $ABCD$, $ABC\delta$, etc. are determined as follows. The chance of a word entered on x counters not falling into any one of k named sectors is given by formula (16). If there are f_x of such words, the expected number of them not falling into any one of the k named sectors is given by (17), and the entire number of counters not falling into any one of the k named sectors is the sum of such expressions for all values of x from unity to the maximum value in the frequency distribution, that is to say, the frequency of negative classes of the *k-th* order, as given by (18).

The arithmetic is most briefly and simply carried out directly as indicated by this equation. With $r=4$, as in our experimental material, we evaluate three frequencies

$$(\alpha) = S\{f_x(\tfrac{3}{4})^x\},\ (\alpha\beta) = S\{f_x(\tfrac{1}{2})^x\},\ (\alpha\beta\gamma) = S\{f_x(\tfrac{1}{4})^x\} \qquad (19)$$

which actually give us fourteen frequencies since $(\alpha) = (\beta) = (\gamma) = (\delta)$, and so forth. We know the value of N (the total number of nouns in the frequency distribution) and are also given $(\alpha\beta\gamma\delta) = 0$, and can then calculate the complete set of fourth-order frequencies, or the frequencies of the *positive* classes or any other set we like. But simply for the sake of more clearly exhibiting how $(ABCD)$, $(ABC\delta)$, $(AB\gamma\delta)$ and $(A\beta\gamma\delta)$ are built up—and we need only concern ourselves with these four since obviously $(ABC\delta) = (AB\gamma D) = (A\beta CD) = (\alpha BCD)$ and so on—we shall calculate these fourth-order frequencies direct from the frequency distribution. The equations, derived from equations (19), are

$$\begin{aligned}
(ABCD) &= S\left\{f_x\left[1 - 4(\tfrac{3}{4})^x + 6(\tfrac{1}{2})^x - 4(\tfrac{1}{4})^x\right]\right\} \\
(ABC\delta) &= S\left\{f_x\left[(\tfrac{3}{4})^x - 3(\tfrac{1}{2})^x + 3(\tfrac{1}{4})^x\right]\right\} \\
(AB\gamma\delta) &= S\left\{f_x\left[(\tfrac{1}{2})^x - 2(\tfrac{1}{4})^x\right]\right\} \\
(A\beta\gamma\delta) &= S\left\{f_x\,(\tfrac{1}{4})^x\right\}
\end{aligned} \qquad (20)^1$$

For once-words, in particular, the last line of formula (20) gives the expected frequency of such words in each sample as 1/4 of the hapax legomena in the total.

[1] To save computational labour, a numerical table is given as Appendix G from which the probability of a word belonging to specified combinations of parts can be read off directly; multiplication of these probabilities by the appropriate value of f_x gives the random partitioning numbers.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Table 14 gives the results, the figures in each of columns 2-5 being the product of f_x (col. 6) by the numerical coefficients of equations (20).

TABLE 14

Calculation of the expected class frequencies (ABCD), (ABCδ), (ABγδ), (Aβγδ) by random partitioning from the combined word count for the samples from Macaulay's essay on Bacon (Yule, 1944)

1 σ	2 (*ABCD*)	3 (*ABCδ*)	4 (*ABγδ*)	5 (*Aβγδ*)	6 Total distribution
1	—	—	—	247.50	990.00
2	—	—	45.87	22.94	366.98
3	—	16.22	16.22	2.70	173.00
4	10.50	15.75	6.13	.44	112.04
5	16.87	10.55	2.11	.07	72.01
6	17.90	6.20	.71	.01	47.00
7	21.02	4.52	.32	.00	41.02
8	19.31	2.74	.12	.00	30.99
9	24.19	2.35	.07	.00	34.01
10	13.27	.91	.02	.00	17.03
11	20.02	.98	.01	.00	24.00
12	16.62	.59	.00	.00	18.98
13	9.06	.23	.00	.00	9.98
14	9.29	.18	.00	.00	10.01
15	12.31	.17	.00	.00	12.99
16	2.88	.03	.00	.00	3.00
17	9.70	.07	.00	.00	9.98
18	6.84	.04	.00	.00	7.00
19	5.90	.03	.00	.00	6.02
20	4.94	.02	.00	.00	5.02
21	0.99	.00	.00	.00	0.99
22	3.97	.01	.00	.00	4.01
23	6.97	.01	.00	.00	7.01
24	1.99	.00	.00	.00	1.99
25	1.00	.00	.00	.00	1.00
26	4.99	.00	.00	.00	4.99
27	2.99	.00	.00	.00	2.99
28	3.99	.00	.00	.00	3.99
29	1.00	.00	.00	.00	1.00
30 up	29.00	.00	.00	.00	29.00
Total	277.51	61.60	71.58	273.66	2048.03

In the same way are the theoretical frequencies calculated for the series of samples from four essays by Macaulay and for that from four works by Bunyan. The calculated frequencies are entered under the heading "Expected" in columns 3, 5, 7 Table 11. There is, on the whole, good agreement between "observed" and "expected" frequen-

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

cies. To exhibit that agreement clearly, we calculate now the percentages of Table 13 from the expected frequencies:

(1) *Macaulay's essay on Bacon*

$$100\,[(ABCD) + (A\beta\gamma\delta)]\,/\,(A) = 100\,(277.51 + 273.66)\,/\,950.71 = 58.1\,\%.$$

The same percentage applies to the other three samples, *B*, *C*, *D*. The calculated percentage is close enough to the average of the observed figures: 59.3%.

(2) *Four essays by Macaulay*

For each sample, *A*, *B*, *C* we have

$$100(607.25 + 319.12)\,/\,1603.05 = 57.8\,\%$$

as against an observed average of 54.1%.

For sample *D* we have

$$100(607.25 + 708.72)\,/\,2330.99 = 56.5\,\%$$

as against the observed 54.3%.

(3) *Four works by Bunyan*

For each sample *A*, *B*, *C*, *D* we have

$$100(449.17 + 259.53)\,/\,1163.68 = 60.9\,\%$$

as against the observed average of 54.4%.

3.4 HOW REAL A PHENOMENON, OR ONLY APPARENT, IS THE HAPAX LEGOMENON?[2]

The preceding analysis enables us to give an objective answer to that question.

For the hapax legomena problem, we consider the number of once-words in connection with that of words occurring in one sample only. Looking at Table 14, we find that 247.5 of the 273.66 words which are peculiar to a sample, that is 90.5%, are accounted for by the hapax legomena in the total.

A word occurring once only in the total can only occur in one sample, and must thus of necessity belong to the class of the form $A\beta\gamma\delta$; moreover, it must also be a sample hapax legomenon. The reverse is not true; a word occurring in the number $(A\beta\gamma\delta)$ need not be a once-word in either total, or sample. The proportion of the number $(A\beta\gamma\delta)$ which is accounted for by hapax legomena in the total consists therefore of hapax legomena in the sample; these are "real" in the sense that they are also hapax legomena in the total.

[2] This is how Whatmough (1956) formulated the problem. By 'real' we mean that uniqueness of a word in the sample reflects uniqueness of that word in the language or text or group of texts, or whatever the universe of discourse.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The relevant quantities for the three series are listed in Table 15. The numbers of one-sample words in each series are the sums of the last four entries in columns 3, 5, 7 Table 11, for "Expected", and in columns 4, 6, 8 for "Observed" frequencies. The observed number of once-words in each series are 990 for the 4 samples from Macaulay's essay on *Bacon*, 1460 for the samples from four essays by Macaulay, and 921 for the samples from 4 works by Bunyan (Yule's tables 4.4., 6.4., 6.7). Since according to formula (20) the expected number of once-words in a sample is ¼ of those in the total, these figures represent also the expected number of once-words in the three series.

TABLE 15

	Observed			Expected		
	Once-words	One-sample words	Percentage	Once-words in the total	One-sample words in *A, B, C & D*	Percentage
4 samples from essay on Bacon	990	1135	87.2%	990	1094	90.5%
4 essays by Macaulay	1460	1666	87.3%	1460	1898	76.9%
4 works by Bunyan	931	1250	74.4%	931	1038	89.7%

In the four sample series from the *Essay on Bacon* we find that the 990 hapax legomena in the total distribute themselves so as to account for 87.2% of the one-sample words, 268 + 280 + 275 + 312 = 1135. This means that up to about 90% the ($A\beta\gamma\delta$) numbers consist of hapax legomena in the sample which are also hapax legomena in the total, and therefore "real". This proportion is sufficiently close to that to be expected under the assumption of random partitioning of nouns, viz. 90.5%, to justify the conclusion that the distribution of the 990 hapax legomena among the four samples is, by and large, the result of random partitioning. Similarly, we find that the one-sample words in the samples from four essays by Macaulay are to 87.3% accounted for by hapax legomena in the total which, of necessity, are also hapax legomena in the samples, as against an expectation of 76.9% on random partitioning. For the four works by Bunyan we find the hapax legomena part of the ($A\beta\gamma\delta$) group to be 74.4% as against an expected 89.7%.

There is a slight increase in the difference between observed and expected magnitude of the hapax legomena part of the one-sample words as we proceed from the samples of one essay by Macaulay to samples of four of his essays, and from there to samples of four different works by Bunyan, as one might expect. But, by and large, the conclusion holds that the hapax legomena in the total account for a great part of the hapax legomena in the sample, namely as far as the latter are also one-sample words;

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

moreover, the occurrence of hapax legomena in the partial distributions can be accounted for to that extent by a statistical mechanism, and thus by chance.

However, just as a one-sample word need not be hapax legomenon in the total, so a sample hapax legomenon need not be a one-sample word: it may be a once-word in more than one sample. The percentage of *all* sample hapax legomena which is accounted for by those in the total results for our three series as follows:

Samples from essay on Bacon: 990/2296 = 43.2%
Samples from 4 essays by Macaulay: 1460/3427 = 42.6%
Samples from 4 works by Bunyan: 931/2084 = 44.7%

To sum up in terms of the question put at the beginning of this section:

1. The hapax legomena in a literary text play a characteristic role in the vocabulary structure of text samples. The group of words peculiar to a sample is in our material made up to something like 70 to 90% by hapax legomena in the total text. That part of the one-sample word group represents also sample hapax legomena. The latter are thus seen to be partly identical with the hapax legomena in the total, and in so far "real", and partly words with other frequency of occurrence in the total text.

2. The extent to which the hapax legomena in the total account for words peculiar to one sample, the $A\beta\gamma\delta$ class of word, and the extent to which the hapax legomena in the sample are "real" depends upon

a. the number and size of the samples;

b. the homogeneity of the distributions, whether they are samples from the same or different texts, the same or different writers, etc.

3. The observed percentages of one-sample words ($A\beta\gamma\delta$), which are hapax legomena in the total—and thus of necessity also in the sample—agree so closely with those calculated by the statistical mechanism of random partitioning, that we may regard the occurrence of hapax legomena among the one-sample words as the result of chance, i.e. of random partitioning of the hapax legomena in the total among the samples.

3.5 RICHNESS (EXTENT) OF VOCABULARY

Instead of concentrating upon the general characteristics of vocabulary distributions, as in the preceding sections, we might wish to stress those features of the distribution which differ from writer to writer, and are thus due to individual style. An important step in that direction is Guiraud's *Les caractères statistiques du vocabulaire* (1953).

Very careful considerations about the true nature of frequency distributions of words resulted in the author proposing as an empirical formula for the extent of vocabulary which is available in the writer's mind, as against the number of words actually used, the expression

$$L = 2N\ (2V_1/V^a) \qquad (21)$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

where L = the extent of vocabulary in the writer's mind
N = the number of mots forts, i.e. occurrences of words other than grammar words
V = the number of different words used, the vocabulary
V_1 = once-words
α = an index of structure, i.e. of distribution of occurrences among vocabulary, calculated on the basis of the following formula:

$$\left[\frac{V_1 V'}{V V_1'}\right]^{\alpha} = N'/N \tag{22}$$

where V_1', V', N' are the counterparts of V_1, V, N in a second sample from the same writer.

In addition, he proposes formulae for the concentration and dispersion of vocabulary.

In this way a most sensible statistical characterisation of style, using the criterion of frequency, is achieved and applied to a series of French poets.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

IV

THE RELATIVITY OF VOCABULARY RATIOS

Descriptive linguistics often requires the use of quantitative methods, that is, of statistics. But it may truly be said, paraphrasing the words of Nurse Cavell about patriotism not being enough, that statistics is not enough. The question put some time ago by Prof. Hönigswald, "How unstatistical is linguistics?" is justified if by "statistics" we understand only the conventional methods as applied in economics, science etc. It is very important to realise that language statistics are often very different, and even paradoxically so, from other types of statistics. In other words: the field of language requires its own statistical methods.

One of the foundations of statistical theory is the Law of Great Numbers. Its import is that characteristic "statistics", like the percentage of a mass of elements with a certain attribute, or the average (mean) of a number of measurements, become more stable, i.e. their errors decrease the greater the number of specimens in the sample. Although it would not, and could not, be correct to say that the law of Great Numbers did not apply to language data, yet it is, at least, much less obvious there, and might, if depended upon indescriminately, lead to very wrong results. For instance, the proportion of Romance words in all works by Chaucer is over 50%, yet we would be very much mistaken if we concluded that this was an estimate of the Romance percentage of vocabulary in individual works by Chaucer. On the contrary, *no* such work has a percentage of Romance words as high as that for the total; moreover, instead of fluctuating around that percentage, with plus-minus errors decreasing as the sample size increases, it rises systematically with text length.

Similarly, if word counts are made, and the words arranged according to their use, that is the times of occurrence, the average number of occurrence per word, and other statistics, increase systematically with sample size.

This shows that linguistics are indeed "unstatistical", if the statistical methods are transferred wholesale from other fields. As we shall show, linguistics has to some extent its own statistics, and thus far linguistics is highly statistical.

4.1 THE DEPENDENCE OF THE ROMANCE VOCABULARY RATIO UPON TEXT LENGTH

It was established in Part I of Volume 1, Sect. 2.5.1, that the percentage of Romance

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

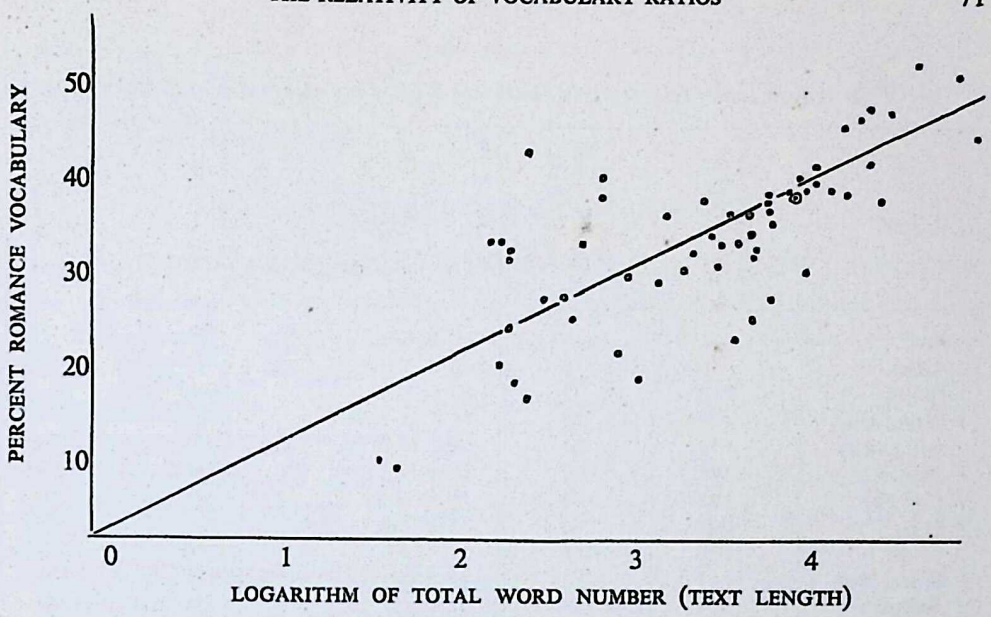


FIG. 6

words in 61 works by Chaucer (Table 3, Vol. I) changes linearly with the logarithm of the text length according to the formula

$$P_v = 10 \log_{10} N \tag{23}$$

where P_v is the percentage of Romance words in the text and N the text length in terms of the number of words; $\log_{10}$ is the common logarithm (Fig. 6).

This result is of a more general interest. If the use of Romance words by Chaucer is considered not as a characteristic of his individual style, but rather as a general characteristic of the English of his time, it would throw light, and very unexpected light, on the mechanism of introducing Romance words. The remarkable fact that the percentage of Romance admixture is a logarithmic function of the text makes the gradual introduction of the foreign element appear to be in the nature of what is known as relative growth. This means that the increase in % Romance words in a small time interval is proportional to the text length at that time.

It follows that the dependence of the percentage of Romance vocabulary upon the length of text, or the fact that the latter is a contributory cause for the former, must be taken into consideration whenever the percentage of Romance words, and probably also the percentage formed by other parts of the vocabulary, is used as an argument for the chronological arrangement of Chaucer's works, or anybody else's. Before a difference in the percentage of Romance words in two works is taken as signifying a chronological difference between them, we must first make sure that the difference is not a merely statistical one, due to the quantitative dependence of the percentage of Romance vocabulary upon the length of text.

डॉ० विद्या निवास मिश्र
प्राध्यापक, भाषा विज्ञान
निजी पुस्तकालय
पुस्तक संख्या.....

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 16

Text length and vocabulary in works of Chaucer (after Mersand 1939)

Work	Total words used (text length)	Number of different words (vocabulary)	Work	Total words used (text length)	Number of different words (vocabulary)
1 Prov.	44	35	31 B. Ship.	3,474	696
2 Adam.	51	38	32 C. Pard.	3,721	844
3 Gentil.	157	72	33 G. SN	4,069	827
4 Wom. Unc.	173	89	34 L.G.W.A.	4,209	489
5 Bal. Comp.	173	84	35 L.G.W.B.	4,230	644
6 Rosem.	183	92	36 Host	4,251	829
7 Lak. St.	195	102	37 D. Sum.	4,571	960
8 Purs.	205	99	38 B. NP.	4,821	984
9 Truth	218	116	39 A. Mil.	5,140	1,016
10 Wom. Nob.	225	113	40 F. Squire	5,181	979
11 Bukton	264	123	41 PF.	5,376	1,101
12 Merc. B.	308	95	42 B. Monk	5,889	1,064
13 Scog.	393	182	43 G. CY.	5,952	1,053
14 A. Ck.	451	193	44 D. WB. Prol.	6,631	1,110
15 Form. A.	473	227	45 A. Prol.	6,683	1,449
16 Fort.	576	251	46 F. Frank.	6,878	1,059
17 Venus	592	234	47 B. ML.	7,852	1,242
18 Am. Comp.	744	246	48 BD.	8,590	1,153
19 Comp. Pite	911	336	49 E. Clerk	8,994	1,231
20 Comp. L.	1,018	304	50 E. Mch.	9,112	1,335
21 B. Thop.	1,156	399	51 R.R.A.	10,583	1,591
22 A.B.C.	1,411	445	52 R.R.C.	11,787	1,597
23 B. Prior.	1,824	401	53 HF.	12,969	1,654
24 H. Manc.	1,976	512	54 Astr.	14,779	884
25 C. Phis.	2,179	562	55 B. Mel.	16,826	1,345
26 Mars.	2,283	618	56 A. Kn.	17,149	2,059
27 Anel.	2,696	675	57 L.G.W.	21,238	1,671
28 D. Friar	2,832	632	58 R.R.B.	21,613	2,267
29 D. WB.	3,210	628	59 I. Pars.	29,672	2,200
30 A. Rv.	3,308	751	60 Boeth.	51,187	2,698
			61 T.C.	64,908	3,032

If an area is fenced off on each side of the straight line at a distance of about 3 % (the standard error of the trend being little over 1 %), this would comprise chance variations in the estimated percentage of Romance vocabulary due to text length. The variation in Romance vocabulary in the 61 works by Chaucer is thus partly explained by difference in text length, namely insofar as the points lie on the straight line, or not farther from it than the control limits at $\pm$ three standard errors of the trend. The points remaining outside that area would represent the works where the error of the estimated value is beyond what might be expected on pure chance, and could, therefore, be

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

considered systematic, for instance, as being due to the subject of the text, to a hypothetical preference for the use of Romance words by Chaucer as he grew older, to the rhyme scheme etc.

4.2 APPLICATION TO DATING OF LITERARY TEXTS

On the basis of the statistical relation established in Formula 23, Fig. 6, it is possible to examine such hypotheses and to arrive at a much better founded conclusion than without the statistical analysis. Let us examine, in the light of these results, Mersand's hypothesis, according to which Chaucer showed a preference for the use of Romance words in his later years, especially when his position at the Court became more established, and more important for him.

Table 17 is a list of works by Chaucer with dates as conjectured by various authors, among them Mersand, who uses his hypothesis of the Romance percentage increasing with length of Chaucer's life (Mersand 1939).

Inspection of the Table, especially of column 4, in conjunction with column 2, does not reveal a very striking trend of the percentage P_v with time, and this seems confirmed by calculating the average of P_v for arbitrarily chosen groups, say, up to 1380, between

TABLE 17

Work	Year	Text Length	Percentage Romance Vocabulary P_v	P_v—10 log N
B.D.	1370	8590	28.79	— 10.55
Comp. Pite	1372	911	28.57	— 1.03
P.F.	1377	5376	34.42	— 2.88
H.F.	1379	12969	37.24	— 3.89
Anel.	1380	2696	29.63	— 4.67
Boeth.	1381	51187	49.85	+ 2.76
T.C.	1384-5	64908	43.47	— 4.65
L.G.W.	1386	21238	36.92	— 6.33
L.G.W.B.	1386	4230	23.91	— 12.35
A. Prol.	1387	6683	39.06	+ 0.81
A.Rv.	1387+	3308	21.71	— 13.49
A. Mil.	1387	5140	26.67	— 10.44
E. Clerk	1387	8994	40.29	+ 0.76
B. Mel.	1387	16826	46.47	+ 4.21
A. Kn.	1387	17149	40.41	— 1.93
I. Pars.	1387	29672	51.27	+ 6.51
Form. A.	1390-3	473	31.72	+ 4.98
Astr.	1391	14779	45.25	+ 3.55
Rosem.	1391-4	183	30.43	+ 7.80
Venus	1393	592	37.61	+ 9.89
Bukton.	1393-6	264	15.45	— 8.77
Fort.	1393-6	576	39.04	+ 11.44
Purs.	1399	205	23.23	+ 0.11

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

1381 and 1391 and after 1391, which averages result repectively as 31.7%, 38.38% and 27.15%. Thus on the face of it and without further statistical refinement, Mersand's hypothesis does not seem to get much support from his own data. However, if the established fact of a concomitant rise of P_v with length of text is taken into account, we must no longer use the absolute differences between P_v as indicating chronological differences, but only what remains of the difference of P_v if the part due to the influence of text length has been deducted. In other words, we must consider the *distances* of the points from the regression line, or, what is the same, the *differences between the observed P_v and that estimated by equation* (23). These differences are given in column 5 of the above table and are plotted against the time co-ordinate as given in column 2 in the following diagram (Fig. 7).

We now get a rather different picture. There is quite obviously a trend, though a slight one, for the works taken into consideration, according to the conjectured time of production. It is specially interesting to note that three out of the five last works, supposedly written between 1391 and 1399, and for which the non-adjusted percentages are, at an average, lower than those for works belonging even to the period 1370-1380, show now appreciable positive differences from the regression line, which points to an assignable cause, in addition to text length.

On the basis of these findings, we can now state that the percentage of Romance vocabulary in works by Chaucer is dependent mainly upon text length, but to a lesser extent, and insofar as it remains unexplained by text length, by a preference for the use of such words increasing as Chaucer grew older.

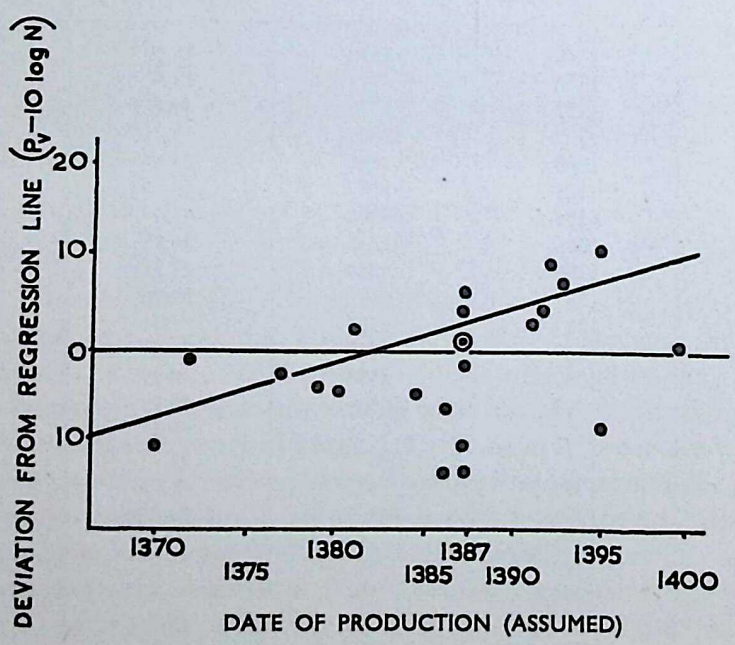


FIG. 7. Deviations from the regression line of Fig. 6 plotted against time.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

4.3 RELATIVE RICHNESS OF VOCABULARY

We now turn to the main object of the investigation, which is to find the explanation for the increase in the percentage of Romance vocabulary with text length, and, in particular, for the linear relation of that percentage to the logarithm of text length.

The method which we shall use is an adaptation of a procedure devised by G. Udny Yule. The problem with which Yule was originally concerned was to compare the vocabulary at the disposal of two authors, and to obtain an answer about the relative richness of vocabulary. An obvious way for arriving at such an answer would be to take text samples of equal length from the writings of each author and count the number of different words occurring in each sample. The quotient of the two figures so obtained would then seem to provide the answer as to relative richness of vocabulary. However, Yule had an idea that the answer would differ according to the sample size, that is the text lengths examined, and thus was not unequivocal. For establishing that hypothesis he uses a certain mathematical function which enabled him to calculate the vocabulary for different text lengths from each author. The ratio of the resulting vocabularies in the two authors then changes actually with sample size, as assumed by Yule from the start.

But Yule's treatment was only theoretical, and he did not produce any evidence from actual word counts that vocabulary ratios actually behave as he predicted they would do.

The use of two types of vocabulary, Romance and Germanic, by Chaucer seemed to provide the required material for testing Yule's procedure. In place of two different authors we have here two different vocabularies, that is blocks of Romance and Germanic words, and, as established empirically, the percentage of Romance words actually changes with text length. If Yule's hypothesis was correct, then his method should provide the possibility for explaining the empirically observed change of Romance vocabulary ratio with text length in Chaucer.

4.4 THEORETICAL DERIVATION OF THE PERCENTAGE ROMANCE VOCABULARY IN WORKS BY CHAUCER

Table 16 gives in cols, 2, 5 the text length in terms of the total number of words, and in cols. 3, 6 the vocabulary, i.e. the number of different words in the text for 61 works by Chaucer. As will be shown, the observed relation between vocabulary ratio and text length arises if the different word groups, Romance and Germanic, *in the total vocabulary at Chaucer's disposal*, are, for texts of different length, drawn upon independently according to a particular law of chance, the Poisson law for the occurrence of rare events. The essence of the method is thus to derive the vocabulary ratio of Romance words in texts of given length from Chaucer's total vocabulary, taking its frequency structure duly into account. We shall speak of that total vocabulary as "la langue" or "la lexique".

Chaucer's total vocabulary is given by Mersand as 8072 words, of which 4189 are of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Romance and 3883 of Germanic origin. Without extensive study of the frequency distribution of Chaucer's vocabulary, nothing definite can be said about its structure. However, we know (Table 20) that the frequency of use of Romance words is, on an average, for texts of any length, about .125, and, consequently, that of Germanic words .875.

If we were to use these two probabilities only, we would be guilty of an over-simplification, which could not but affect the results. We would imply that the probability of occurrence was the same for all Romance words, and, similarly, another probability the same for all Germanic words, which is obviously not true. But let us see whether the available information would enable us to arrive at a more detailed structure in reasonable harmony with general principles. We know that only a part of Chaucer's total vocabulary was used in the 61 texts considered, since even the longest, *Troilus and Criseyde*, with over 64,000 words has only about 3,000 different words, as against the over 8,000 of Chaucer's total vocabulary. This does not mean that about 5,000 words were not used at all, since the vocabulary of the smaller texts need not always belong to the 3,000 words of the longest. But it would seem sensible to assume that about 5,000 words were drawn upon only very rarely. We shall, therefore, let them account for only a very small percentage of the total occurrences, say, .005 of each, Romance and Germanic words. Of the 3,000 words which we regard as the more special vocabulary on which Chaucer drew for the 61 works under consideration, about 1,300 are Romance and 1,700 are Germanic. In each group, we shall make a small proportion of words account for the greater part of occurrences, 100 G words for .60 (this part containing most of the grammar forms and other words with very high frequencies) and 300 R words for .10, leaving 1,600 G words to account for .27 and 1,000 R words to account for .02 of the occurrences. The complete scheme is exhibited in table 18.

For each group, the number of different words, V, in texts of different length, N, is calculated according to the Poisson law of rare events as

$$V = L(1 - e^{-x}), \tag{24}$$

where $x = Np/L$,
and $e^{-x} = 1/e^{x}$,

which values are given in tables of the exponential function e^{x}.

The argument implied in using the above formula for V is as follows.

According to the law of rare events, the probability of one of L words not occurring in a sample of 1 word is $e^{-1/L}$; the probability of one of L words not occurring in a proportion Np of a text of length N is $e^{-Np/L}$, and consequently the probability of its occurrence is $1 - e^{-Np/L}$. The probable number of words of the group in a text of length N is then $L(1 - e^{-Np/L})$.

The following table exhibits the results of the calculation and gives in rows 4 and 8 the G and R words respectively, and in rows 9, 10 the total vocabulary and the percentage of R words respectively.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 18

Group	No. of words (L) in groups of Chaucer's total vocabulary	Proportion (p) of text length accounted for by L
1	100 } 3883 Germanic words	.60
2	1600 } 3883 Germanic words	.27
3	2183 } 3883 Germanic words	.005
4	300 } 4189 Romance words	.10
5	1000 } 4189 Romance words	.02
6	2889 } 4189 Romance words	.005

TABLE 19

Sample vocabulary calculated by compound Poisson law

Total Vocabulary ("lexique") Word Groups	Text Length (N) 500	1000	2000	4000	8000	16000	32000	64000
	V_t	V_t	V_t	V_t	V_t	V_t	V_t	V_t
1 .... 100 } 3883 G	95	99	100	100	100	100	100	100
2 .... 1600 } 3883 G	131.2	253	459	785.6	1185.6	1493	1593	1600
3 .... 2183 } 3883 G	2.2	4.4	8.7	17.5	35	70	140	262
Sum G words	228.4	356.4	667.7	903.1	1320.6	1663.0	1833.0	1962.0
4 .... 300 } 4189 R	46.2	85.8	146	220.5	279.3	298.6	300	300
5 .... 1000 } 4189 R	10	20	39	77	148	274	473	7022
6 .... 2889 } 4189 R	2.6	4.9	9.8	19.6	37.6	78	156	303.3
Sum R words	58.8	110.7	194.8	317.1	464.9	650.6	929.0	1325.3
Total Vocabulary V ($G + R$)	287.2	467.1	762.5	1120.2	1785.5	2313.6	2762.0	3287.3
% R words	20.5	23.6	25.6	26.0	26.0	28.1	33.6	40.3
Δ % R words	3.1	2.0	0.4	0	2.1	5.5	6.7	
Voc./Text length V/N	.58	.47	.38	.30	.22	.15	.09	.05

Figure 8 is a bilogarithmic graph of the relation between the vocabulary V and the text length N. The continuous curve represents the theoretical values for V (Row 9, Table 19) plotted against text length, the points represent the observed values in 61 texts by Chaucer (Cols. 2 and 5, Table 16).

4.5 DISCUSSION OF RESULTS

1. The observed and theoretical values of the vocabulary show very similar trends with text length. The theoretical curve may be considered a satisfactory fit of the observed

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

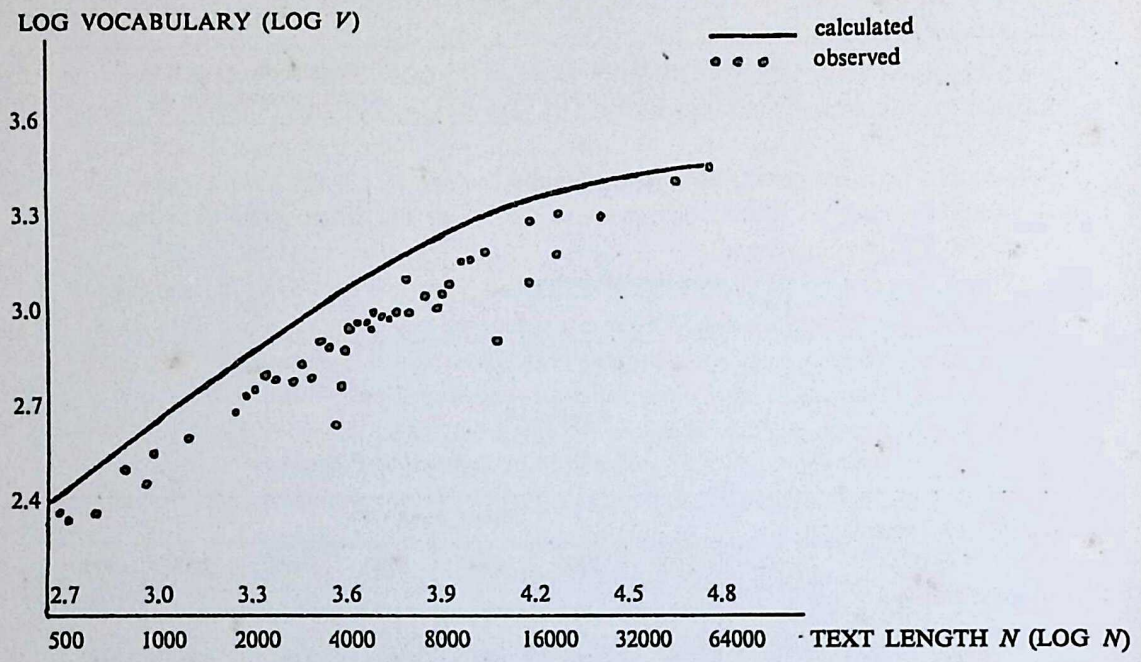


FIG. 8

scatter of points. The trend is fairly linear up to texts of 8,000 words, and from then on curves towards the saturation point. Inspection of the table reveals very clearly the way in which the vocabulary grows with text length.

A straight line trend on a bilogarithmic grid implies a uniform increase or decrease of the ratio between the variables, according to whether its angle with the horizontal is greater or smaller than 45°; in our case, it implies that the ratio V/N is uniformly decreasing between $N = 500$ and $N = 8000$ approx. A curved trend on such a grid implies an accelerated or decelerated change of the ratio between the variables, according to whether the bend is upwards or downwards; in our case, the bend from $N = 8000$ on means a deceleration in the decrease of V/N. These results are quite general, and if supplemented by the information about the first part of the curve—not shown on our graph—bending upwards and indicating an accelerated decrease of the ratio V/N, give this picture of the changing relation between vocabulary and text length: starting with a very high value close to unity, the ratio V/N decreases, first rapidly, then slower until it settles down to a uniform rate of decrease persisting over a certain medium range, in our case for text length from $N = 500$ to $N = 8000$; from then on, the ratio V/N decreases slower and slower, as N increases. (Fig. 8 and last line of Table 19.)

The six groups of words according to frequency of occurrence, grow at characteristically different rates. Group 1 grows fastest, reaching completion in texts of only 2,000 words; next come groups 2 and 4, which reach completion in texts of 64,000

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

words; group 5 attains even in the longest texts only a little over $^2/_3$ of its size; and groups 3 and 6 little more than $^1/_{10}$.

2. The vocabulary ratio increases with text length N, from 20% to 40%, and insofar the mechanism of the observed behaviour of the vocabulary ratio is explained as a consequence of the aleatoric nature of vocabulary size in these texts, or as a stochastic function of the extent and structure of "la lexique"; that is, texts of different length can be regarded as random samples of "la lexique" as the parent distribution, as far as the size of vocabulary is concerned.

3. We come now to the second part of the problem as put on p. 75: the linear relation between vocabulary ratio and log text length.

In two pilot fittings, two and four components were used, as against the six components used above, according to the following schemes (Table 20).

The ratios and percentage differences obtained are listed in table 21.

TABLE 20

Group	L	p	Group	L	p
G	1700	.875	G 1	100	.600
R	1300	.125	G 2	3783	.275
			R 3	1000	.100
			R 4	3189	.025

TABLE 21

N	2 components %R	Δ%	4 components %R	Δ%
500	12		21	
		3.0		3.0
1000	15		24	
		3.0		3.0
2000	18		27	
		4.0		2.0
4000	22		29	
		7.4		.3
8000	29.4		29.3	
		8.0		.5
16000	37.4		29.8	
		4.6		.2
32000	42		30	
		0.2		6.0
64000	43.2		36	

For text lengths up to 4000 words, we find that as text length doubles, and its logarithm, therefore, increases in steps of .3 approx., the vocabulary ratio increases by sensibly equal steps of 3%. Thus equal differences in vocabulary ratio correspond to

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

equal ratios of text length. The regression coefficient is for this part of the curve very nearly $b = 3/.3 = 10$, the same as found empirically (formula 23).

This relation is not so clearly shown by the ratio for the 6-component fit (line 11, table 19). The explanation is this. Increasing the number of components with a view to fitting the distribution of total vocabulary according to text length by a compound Poisson curve is a double-edged sword. Although we may succeed in improving the fit of the *total* vocabulary as intended, it may impair the relation of the vocabulary *ratio* to text length, and this may account for the poorer fit of the vocabulary ratio-log N relation.

As N grows beyond 4000, the vocabulary ratio grows faster than log N, but after 20,000 bends towards the maximum. However, even so, the average vocabulary ratio difference for the 6-component fit over the total range from $N = 500$ to $N = 64000$ is 2.82 and thus close enough to the theoretical 3.0.[1]

We can thus say confidently that vocabulary ratios change with text length, and that to some extent, and differentiating carefully between the different probabilities of occurrence of the vocabulary components, the changes are such as to be compatible with the assumption of the texts being random samples of the "lexique" of specified extent and structure. We have also accounted for the relation between vocabulary ratio and logarithmic text length, with the qualifications given above. It follows in our case from the same proposition which enabled us to calculate the ratios as increasing with text length.

Since the mathematical method for calculating the size of vocabulary components on the basis of their distribution in "la lexique" is perfectly general, it follows that by "component" we may understand any part of the vocabulary whose frequency of occurrence differs significantly from the rest. We may, for instance, note in literary texts other characteristics besides words—particular constructions, idioms, similes and metaphors, unusual grammatical forms, such as unusual declination of a noun or unusual conjugation of a verb, peculiarities of word order etc.—and inquire about their ratio to the usual or conventional form in texts of given length, as different from their overall ratio.

One might be tempted to generalise also the second finding, and say that the vocabulary ratios generally change proportionally with the logarithm of text length, but, I am afraid, this may not be justified on the slender evidence before us. It would, however, be most interesting to explore the matter, and see whether more evidence could be found, for either an unqualified or a qualified confirmation of that conclusion.

4.6 THE RELATIVITY OF VOCABULARY RATIOS

This investigation establishes and explains what I have called in the chapter heading "the relativity of vocabulary ratios".

[1] Better agreement between observed and theoretical values could hardly be expected since the assumed frequency distributions are only rough approximations of the true vocabulary structure.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The principle of relativity as formulated by Einstein demands that the formulation of the laws of nature or the description of events in nature should be "*invariant*", that is, independent of the reference system. This is in the Theory of Relativity accomplished by certain changes in the conventional transformation formulae so as to conceive of the description of the event with regard to different systems of reference as transformations of one another.

Applying these ideas to the statements of descriptive linguistics of the kind called vocabulary ratios, and more particularly to the one in which we are here interested, viz. the Romance vocabulary in works of Chaucer, we were at the outset confronted with apparently contradictory statements of that ratio in different texts. However, by conceiving of the vocabulary ratio in the various texts as the realisation in "la parole", of the basic relation of Romance to total vocabulary at Chaucer's disposal, or of that ratio in his "lexique", we have succeeded in recognising the essential agreement between all the different statements about the ratio: they all express the same event or fact, namely, Chaucer's "lexique", or his total vocabulary of *R*- and *G*-words, in extent and structure, the difference between them being due to different text length. Accordingly, they appear now as transformations of one another, the transformation formula being the regression equation

$$P_v = C + 10 \log_{10} N$$

A change in text length (N) in the formula has the effect of changing the vocabulary ratio in the way demanded by theory, that is, through the influence of text length as sample size upon the basic "lexique".

4.7 CAUSE AND EFFECT IN LANGUAGE

These results are of particular interest in connection with the controversy of determinism vs. indeterminism in language.

The deterministic view regards linguistic expression as the result of a deliberate choice of linguistic units on the part of the speaker. Thus, A. C. S. Ross thinks that if he says "teapot", it is because he just *wants* to say this, and believes he has thus disposed of the hypothesis of chance as a factor in language.

Apart from misunderstanding the statistical hypothesis (Sect. 1.7), Ross's objection misses the point that the views as to what requires a "cause" have changed in the course of the history of science, and that it is not just commonsense which decides what needs a cause, and what could happen without one, but which assumption was more effective in establishing relations between phenomena. "If a bullet goes on moving, is a cause necessary to keep it going, or would a cause be necessary to stop it? . . . . The *Ancients* thought that in the absence of a cause it would be at rest, and that a cause was necessary to keep it moving (just as Prof. Ross thinks that in the absence of a cause or motive there would be no occurrence of the word 'teapot', and a cause was needed for its appearance). *Newton* thought that in the absence of a cause

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

it would go on moving with uniform velocity, and that a cause was necessary to change its velocity. *Einstein* thought that in the absence of a cause it will move with uniform acceleration, and that a cause is necessary to change the acceleration. *We choose the assumption which subsequently provides the best account for the relations between phenomena*." (Einstein in Samuel, 1952).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

V

THE CHARACTERISTIC CONSTANT OF THE TYPE-TOKEN RELATION

5.1 DIFFERENT INTERPRETATIONS OF THE CHARACTERISTIC CONSTANT *K*

The Characteristic *K* has been interpreted in various ways. Yule, who devised it as a statistical constant of word frequency distributions regarded it as a measure of concentration of vocabulary. Whereas the simple type-token ratio changes with size of text sample, and cannot, therefore, be used as a constant text characteristic, *K* expresses the type-token relation independent of sample size, or, in other words, *K* is a statistical constant. Williams (1946) who applied it to the distribution of animal species according to frequency of occurrence called its reciprocal a measure of diversity. In a way, the two interpretations are very much the same. They are also similar in that they use certain specific distributions for deriving it: Yule uses the properties of the Poisson law, and Williams uses the logarithmic series. Contrary to this, the derivations by Good and Herdan do not use an hypothesis about the statistical population. According to Good (1953) *K* can be conceived to be the repeat rate of species, words etc. in a great number of occurrences of such events, i.e. the probability that words, say, chosen at random from a text will be the same dictionary word; according to Herdan (1955 and Vol. 1, Sect. 3.5) *K* can be described in great samples as the coefficient (squared) of variation of the mean number of occurrences per word.

Moreover, I have shown that *K* is more than a stylostatistical parameter; it appears to characterise the relation between vocabulary and occurrence at certain stages in the development of language, thus epitomising the "langue-parole" relation. This means that I have accorded to it the role of a general characteristic of language structure, instead of regarding it only as a feature of individual style.

5.2 THE GEOMETRICAL INTERPRETATION OF THE REPEAT RATE

1. From a frequency distribution such as given below for the alphabetic symbols of written Dutch (table 22), we calculate the repeat rate as

$$\varkappa_p = \sum_{1}^{26} p_i^2 = 0.0827 \qquad (25)$$

In general, we may represent a printed language having n different alphabetic symbols as a point *P* in *n*-dimensional space with probabilities p_i as co-ordinates. Then $\sum_{1}^{n} p_i^2$ represents the square of the distance of the point *P* from the origin. That point

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 22

Frequency distribution of alphabetic symbols in written Dutch[1]

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*
n_1:	65	17	13	61	200	6	37	30	66	17	23	35	22
	N	*O*	*P*	*Q*	*R*	*S*	*T*	*U*	*V*	*W*	*X*	*Y*	*Z*
n_1:	108	57	13	0	63	35	58	18	24	18	0	0	14

$n_1 = 1000\, p_i$.

lies also in the hyperplane $\sum_1^n p_i = 1$.

It follows from the definition that the repeat rate is not uniquely specified. All points which lie on the intersection of the (n—1)-dimensional sphere with $r^2 = \sum_1^n p_i^2$ and the plane $\sum_1^n p_i = 1$ have the same distance from the origin.

This means that the parameter would be the same for any language in which the frequency distribution of alphabetic symbols was only a permutation of the original one, i.e. any language with identically the same frequencies, but attached to different letters.

A sufficient estimate of $\varkappa_p$ is given by the statistical quantity

$$K = \Sigma \frac{n_i(n_i - 1)}{N(N-1)} \tag{26}$$

where N = number of letters in the sample,
n_i = frequency of occurrence of the i-th letter.
The proof is given in Appendix D.

For great N this becomes

$$K = \Sigma \frac{n_i^2}{N^2} - \frac{1}{N} \tag{27}$$

which we recognise as Yule's Characteristic.

The variance of K is given by

$$\sigma_\varkappa^2 = \frac{0.0168\, N + 0.1182}{N(N-1)} \tag{28}$$

In the case of equi- or rectangular distribution, $p_i = 1/n$, and we have for the repeat rate

[1] M. de Vries, *Kwantificering in taal en cryptoanalyse* (Rapport SP 32 van de Statistische Afdeling, Mathem. Centrum) (Amsterdam, 1954).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\varkappa_r = \sum_1^n \left(\frac{1}{n}\right)^2 \tag{29}$$

For a language with 26 letters, such as Dutch or English, we get

$$\varkappa_r = \sum_1^{26} \left(\frac{1}{26}\right)^2 = 0.0385$$

The distribution of $\varkappa_r$ approaches that of the non-central Chi-square distribution. In all other cases it is dependent upon the parameter p_t. Sometimes the parameter $\varkappa_p$ is called the Coincidence constant.

2. The geometrical interpretation of the repeat rate as epitomising the use of letters, or phonemes, can be extended to other levels of the language, e.g. if the repeat rate is used for epitomising the type/token relation in a given text or group of texts. This is where that parameter was originally used. In that case the point P in n-dimensional space represents the text, and n is the vocabulary of the text, i.e. the number of different dictionary items with probabilities p_t as co-ordinates, whose estimates are the relative frequencies of the particular words. Then $\sum_1^n p_t^2$ is the square of the distance of P from the origin. That point lies also in the hyperplane $\sum_1^n p_t = 1$.

It follows from the definition of that parameter that the text is not uniquely described by it. The geometrical locus of all points P lie on the intersection of a hypersphere with $r^2 = \sum_1^n p_t^2$ with the hyperplane $\sum_1^n p_t = 1$, all these points having the same distance from the origin, and thus representing the same repeat rate.

This shows very clearly that Yule's constant K as a sufficient estimate of the repeat rate $\sum_1^n p_t^2$ depends only upon the distribution or number of vocabulary items according to frequency of occurrence, and not upon which particular word has which particular frequency. Permutation of the p_t among the vocabulary items of the text would leave K unchanged. However, the purely geometrical interpretation of the repeat rate is in the case of language subject to an important restrictive condition, which will be dealt with in the next section.

To put it succinctly: although a given universe of discourse is, by and large, characterised by the repeat rate, the obverse is not true: identical repeat rates do not necessarily imply the same universe of discourse. It follows from the first part of the above statement that the characteristic K as an estimate of the repeat rate from samples of a given universe of discourse will not differ by more than can be accounted for by fluctuations of sampling; a difference significantly in excess of this would make us suspect that the method of sampling was not such as would conform with the requirements of random sampling.

On the other hand, it follows from the second part of the statement that if it is doubtful whether a number of samples have come from the same or from different universes of discourse, even the sensible equality of the values of the characteristic K

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

calculated from them would not compel us to conclude that the samples have all come from the same universe of discourse, but would point only to a formal similarity in the frequency pattern of vocabulary, and thus to similarity of style.

5.2.1 The Restrictive Condition for the Geometrical Interpretation of the Repeat Rate

So far, we have given the geometrical interpretation of the repeat rate without any restrictive condition. It is, however, of the greatest importance for the sensible application of geometrical methods of language structure to realise that there is a restrictive condition.

The repeat rate for the phoneme (letter) distribution of a particular language is, although not strictly unique, typical for that language and can serve as a parameter of the language in that respect, because it is an empirical fact that the frequencies or probabilities attached to the units in the phonemic (alphabetic) system of different languages do not happen to be just permutations of one another in different languages. It follows that the parameter epitomising the use of phonemes (letters) of a language is, by and large, unique. Since it is a statistical parameter, uniqueness must be interpreted in the statistical sense, namely, as a very high probability.

Similarly, it has been empirically established that the relative frequencies attached to particular words in a literary text are characteristic, to a greater or lesser extent according to the type of words, not only of the particular text, but of the use of the vocabulary items as such. It follows that the repeat rate for a text will, to a great extent, depend upon its particular vocabulary. Since different writers and texts will, in general, differ in their vocabulary, they will be characterised by different repeat rates. It is this empirical fact of the sensible uniqueness of that parameter for the vocabulary comprised in a given text, which justifies its use as a characteristic of style (Sects. 6.1, 11.5, Vol. I).

This is uniqueness—cum grano salis—in the statistical sense that the probability of its arising in the same value for another complex of basic data of the given type, e.g. letters, phonemes, vocabulary, is so negligibly small as to amount to non-existence. This does not mean that it is *impossible* that two languages should differ in the use of the same phonemic system only through a sort of permutation of the probabilities of particular phonemes, or two literary texts should differ only through the relative frequencies of the same vocabulary items being just permutations of one another, or that a very different vocabulary should occur with the same frequency pattern, but the *probability* of such an event arising in reality is so small that it can be neglected. What we call here uniqueness of the repeat rate is, therefore, statistical uniqueness based upon the law of great numbers.

5.3 THE RELATION BETWEEN V_m AND THE REPEAT RATE

The different interpretations of K by Good and the author provide an excellent opportunity for demonstrating the use of dualities in linguistics (See Sect. 17.2).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

It was stressed in Sect. 12.3.1, Vol. 1, that the vocabulary-occurrence relationship represented the fundamental linguistic duality. There are two distributions involved. There is the distribution (as used by Zipf) in which the individual items of vocabulary are listed on the abscissa and the frequency of each is plotted on the ordinate. But the same data can be differently plotted by reversing the roles of vocabulary and frequency. We could list on the abscissa the various frequencies that occur, and plot on the ordinate the number of vocabulary items showing each frequency. This is the distribution with which Yule chiefly worked. The two plots are not contradictory or competing descriptions; rather they are complementary ways of summarising the same data.

In terms of geometry, we have here two figures (probability density curves) F and G, representing the relation between the co-ordinates P and Q (variable and probability), such that when figure F is changed into figure G, the property P is changed into the property Q. The relation thus satisfies the formal requirement of mathematical duality. It also satisfies the intrinsic requirement that the dual aspects of the collective be "contemporaneous and complementary" and not "for or against" one another.

According to which distribution one has in mind, that of word occurrences according to the probability of occurrence of individual words, or that of words according to their frequency of occurrence, K is derived either as a repeat rate of individual words (Good) or as the coefficient of variation (squared) of the mean frequency of occurrence of words (Herdan).

The formula for Yule's Characteristic is

$$K = \frac{S_2 - N}{N^2}$$

$$\text{where } N = \sum_1^s r n_r$$

$$\text{and } S_2 = \sum_1^s r^2 n_r \tag{30}$$

and where n_r is the number of different words occurring with the frequency r, and Σ is the sign for summation.

In formula (30) both r and n_r are the frequencies of word occurrences and vocabulary respectively, and may by division through the corresponding totals be transformed into relative frequencies. For great samples, these relative frequencies are sensibly stable and may be regarded as estimates of the corresponding probabilities, $p_r = r/N$ and $\pi_r = n_r/n$ $(n = n_1 + n_2 + \ldots . n_s)$ respectively.

Using the first transformation, we have

$$K^* = S_2/N^2 = \sum_{r=1}^{s} \frac{r^2}{N^2} n_r = \sum_{r=1}^{s} p_r^2 n_r \tag{31}$$

and summing over the individual p_r instead of over their groups n_r, we may write

$$K^* = \sum_{\mu=1}^{n} p^2 \tag{32}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

which shows K to be the repeat rate of particular words. (For a more satisfactory proof see Appendix D.) Using the second transformation we write

$$K^* - \frac{1}{n} = S_2/N^2 - 1/n = \frac{n \Sigma r^2 n_r - [\Sigma r n_r]^2}{n [\Sigma r n_r]^2} \tag{33}$$

Dividing numerator and denominator by n^3, the expression on the right becomes

$$\frac{\{\Sigma r^2 \pi_r - [\Sigma r \pi_r]^2\}/n}{[\Sigma n \pi_r]^2} = \frac{\sigma_r^2/n}{M_r^2} \tag{34}$$

which shows K (more precisely K^*) for great n, to be equal to the coefficient of variation of the mean of the variable r.

The apparent contradiction between Good's and my own interpretation of the constancy of K is thus resolved by realising that the two interpretations are dual to one another (Sect. 17.1, 17.2). This being so, they partake fully of the characteristics of mathematical duality, viz. of being not rival hypotheses, but complementary and contemporaneous ones. The duality arises out of the texture of linguistic space, and thus reveals more about that space than any single statement could have done. This is the great value of duality in language. Apart from it, its recognition as a duality contributes to disposing of controversies in linguistics. The conception of both views as a duality in the strict sense of the term combines both views into one harmonious whole, free from contradiction.

It is just in keeping with the versatility of this parameter for describing linguistic structure that it should be capable of being defined mathematically as a coefficient of variation, as a repeat rate, and as an expectation, i.e. a mean value, of a probability (See Sect. 14.2).

5.4 THE STABILITY OF V_m IN TRANSLATIONS

In the course of an investigation of quantitative changes in the development of English, J. Krámský compared the frequency distribution of dictionary words in samples from the original text of the Homilies of Aelfric, comprising 9752 occurrences, with that in samples from a translation into modern English (1846), comprising 10902 occurrences.

Table 23 gives the frequency distributions in detail.

We see (last but 3rd col. of statistical parameters) that v_m is sensibly constant, from which we may infer that the type-token relation is essentially the same in original and translation. This is as it should be if the translation into modern English retains the style of the Old English original.

But we may ask, whether the constancy of v_m is a consequence of the preservation of the individual style of the author, or of the English of the period. This is a question of general interest for the interpretation of K.

K by its mathematical structure is a characteristic of the type-token relation in a

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 23

Word count of Aelfric's Homilies (Krámský)

No. of dictionary words in the texts		Frequency of Occurrence	No. of dictionary words in the text		Frequency of Occurrence	No. of dictionary words in the text		Frequency of Occurrence	No. of dictionary words in the text		Frequency of Occurrence
Old	New		Old	New		Old	New		Old	New	
941	950	1	2	6	23	2	1	49	–	1	126
291	289	2	3	2	24	–	2	50	–	1	127
184	128	3	3	–	25	–	3	51	–	1	151
99	102	4	1	5	26	1	–	53	–	1	153
59	66	5	–	1	27	2	1	54	1	–	163
40	42	6	2	–	28	2	2	57	1	–	169
41	32	7	–	1	29	–	1	59	–	1	170
20	29	8	1	1	30	–	3	62	1	–	176
15	19	9	3	–	31	–	1	66	1	–	180
11	9	10	3	1	32	–	1	67	–	1	187
8	13	11	–	1	33	1	3	70	1	–	194
14	9	12	3	1	34	–	1	71	2	–	218
14	8	13	1	1	35	1	–	72	–	1	227
11	11	14	1	3	36	1	–	77	1	–	245
5	7	15	1	1	37	–	1	79	–	1	272
3	9	16	–	4	39	–	1	81	–	1	306
4	5	17	–	1	40	–	1	84	–	1	326
6	3	18	1	–	43	1	–	87	–	1	384
2	2	19	2	–	44	–	1	88	1	–	417
2	2	20	2	–	45	1	–	94	1	1	472
4	2	21	1	1	47	1	–	116	1	–	512
4	4	22	1	–	48	–	1	121	–	1	808

Statistical parameters	$\Sigma f_x X$	No. of dict. units (N)	Mean (X)	$\Sigma f_x X^2$	Varance (σ^2)	Coefficient of Variation (v_x)	v_m	v_m^2	$1/N$	K^*
Old text	9752	1834	5.3172	1,104,434	23.75	4.462	.1042	.0108	.00054	.01161
New text	10902	1807	6.0332	1,647,101	29.563	4.900	.1152	.01322	.00055	.01385

given linguistic code; in our example, of the type-token relation of the word code of a language. However, as language develops, the word-code becomes more and more individualised and in different writers to a different extent. This has as a consequence that the higher degrees of individualisation must be regarded as, to some extent, different codes, which is what we mean by saying that writers differ in style. It explains why K should be a characteristic of language as well as of style.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

5.5 v_m AS CHARACTERISTIC FOR DISTRIBUTIONS OF LINGUISTIC FORMS OTHER THAN WORDS

Murray Fowler used v_m for testing the stability of the type-token relation where "type" stands for phonemes and "token" for their occurrence in continuous pieces of writing. He realised that when the phoneme is the unit, the value of N, as the number of phonemes in a language, remains constant, and with it $\overline{X}$, and that it was therefore, strictly speaking, not necessary to calculate v_m for the sole purpose of obtaining a stable value. However, it seemed advantageous to apply the same method of characterisation in all stages of an investigation aimed at establishing the "langue-parole" relation on the phonemic, morphemic and syntactic level (Fowler, 1957).

Four examples were examined, all chosen from modern English. Each sample was transcribed into a phonemic alphabet of 24 consonants and 6 vowels. Three of the samples consisted of 5,000 consecutive phonemes each; the fourth, a story for children, is complete in 501 phonemes. The samples of 5,000 were each divided into 5 samples of 1000 each, and the mean, standard deviation and coefficient of variation of the sampling distributions, v_m, were calculated for both the total and the partial distributions. The results given in Table 24 show that the type-token relation of phonemes is not significantly disturbed by individual differences in style (author or subject) and may, therefore, be regarded as part of the structure of "la langue".

5.6 THE EFFECT UPON K OF TRUNCATING THE FREQUENCY DISTRIBUTION

The examination of the behaviour of the relative frequencies of dictionary words with increasing text length shows that whereas the p_i of high frequency words settle down to sensibly stable values as the text length becomes great, the p_i of low frequency words continues to decrease with great text length. This results in a slight but systematic decrease of K as the sample size approaches the total (Sect. 20.5).

It has been suggested that this might be used as a short cut for calculating K, and v_m, by obtaining it for a truncated word count. The argument put forward was that the proportion of times that the high frequency word occurs is proportional to the amount of the text sample, and therefore $\Sigma f_x X^2/[\Sigma f_x X]^2$ must be essentially constant, numerator and denominator being dimensionally the same. For instance, we might arbitrarily limit the calculation to words occurring 100 times or more in the total, 50 times or more in samples of about $^1/_2$ the size of the total, and to words occurring 25 times or more in samples of $^1/_4$ the size of the total. This, it was claimed, would accordingly reduce the calculation and give values of K very close to those for the complete word count.

Unfortunately, the argument is fallacious, both from the theoretical and the practical point of view. Theoretically, it follows from the interpretation of K as the repeat rate of dictionary words that a truncated frequency distribution in which only high frequency words are left must give a substantially higher K; from the practical

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

TABLE 24

(After M. Fowler, 1957)

Statistical Constants	*A Drive in the Country* (Graham Greene)						*Fielding and Sterne* (Graham Greene)						*Concepts of the Calculus* (C. B. Boyer)						*The Story of a Fierce Bad Rabbit* (Beatrix Potter)
	1st 1000	2nd 1000	3rd 1000	4th 1000	5th 1000	5000	1st 1000	2nd 1000	3rd 1000	4th 1000	5th 1000	5000	1ts 1000	2nd 1000	3rd 1000	4th 1000	5th 1000	5000	
Sample Size																			501
Mean	33.2	33.2	33.3	33.5	33.1	166.367	33.3	33.2	33.1	33.4	33.2	166.133	33.3	33.2	33.0	33.2	33.0	165.667	16.7
Standard deviation	21.7	23.6	23.9	23.2	24.8	114.041	24.6	25.1	23.8	23.6	25.1	120.704	25.4	25.5	25.3	27.5	25.9	124.865	14.297
v_m	.1191	.1297	.1311	.1396		.125139	.1348	.1382	.1311	.1289	.1381	.13265	.1393	.1403	.1395	.1483	.1432	.137534	.1563*

* It was observed that v_m did not reach a stable value until about 1000 phonemes had been counted. This may explain the deviation of v_m for this sample.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

point of view it would be no great reduction in arithmetical labour because it is not forming the squares of the low frequencies which is troublesome but that of the high frequencies.

To show the correctness of the theoretical objection, the word counts from *The Captain's Daughter*, tables 6-8, Vol. 1, were truncated as follows. The distribution for the whole book (table 6) was truncated so that only words of 50 occurrences and more were left. For the partial distributions of tables 7 and 8 the total numbers of occurrences are .32, .173 and .164 of the total occurrences of table 6; multiplying the truncation value of 50 for the total with these proportions we arrive at 16, 8, and 8 approximately as the truncation values for the distribution of tables 7 and 8. The following table lists the basic quantities and the values of K^* for these distributions.

TABLE 25

	Table 6	Table 7	Table 8	
			(*a*)	(*b*)
$\Sigma f_x X$	12079	4314	2311	2240
Σf_x	70	81	99	82
M_x	172.52	53,25	23.34	27.4
$\Sigma f_x X^2$	4000667.930	526,307	153,705	143.698
K^*	.02742	.02823	.02877	.02684

As predicted we see that K^* is entirely different from K^* for the complete word counts (Table 10, Vol. I), being about 4 times the latter. In terms of K as the repeat rate, this means that whereas for the complete word count of whatever size of sample a pair of words picked at random will be the same dictionary word about once in 150 trials, for the truncated samples it will be the same dictionary word about once in 36 trials.

5.7 THE MATHEMATICAL RELATION BETWEEN GREENBERG'S INDEX OF LINGUISTIC DIVERSITY AND YULE'S CHARACTERISTIC

At the first glance the two measures mentioned in the title are different both in form and meaning: Greenberg's constant is meant to measure the diversity of language in a speech community, and is mathematically the complement to unity of the sum of the squared probabilities of these languages. Yule's constant characterises the vocabulary-occurrence (type-token) relation in a given text, and is mathematically the ratio of the second to the squared first moment of the word count distribution.

5.7.1 Greenberg's constant

The examination of any map of linguistic areas will show regions of greater diversity and others of relative uniformity, while still others may seem intermediate between these extremes. Greenberg's aim is to have quantitative objective measures of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

diversity by which to replace such subjective impressions. The simplest model, called the monolingual non-weighted method A, may be described as follows (Greenberg, 1956).

If from a given area we choose two members of the population at random, the probability that these two individuals speak the same language can be considered a measure of its linguistic uniformity. If everyone speaks the same language, the probability that two such individuals speak the same language is obviously 1, or certainty. If each individual speaks a different language, the probability is zero. Since we are measuring diversity rather than uniformity, this measure must be subtracted from 1, so that our index will vary from 0 indicating the least diversity, to 1 indicating the greatest.

If the area comprises speakers of r languages and the proportion of speakers of language i is p_i ($i = 1,2, \ldots, r$), then the total probability of choosing two speakers of the same language is the sum of the probabilities of such an event for each individual language i.e. $\sum_i p_i^2$. Subtracting this sum from 1 as suggested above, the formula for the coefficient of linguistic diversity becomes

$$A = 1 - \sum_{i=1}^{r}(p_i^2) \tag{35}$$

This may be illustrated by a hypothetical example. If in a population three languages are used and the proportion of speakers is as $\frac{1}{8} : \frac{3}{8} : \frac{1}{2}$, then

$$A = 1 - \{(\tfrac{1}{8})^2 + (\tfrac{3}{8})^2 + (\tfrac{1}{2})^2\} = 1 - \frac{26}{64} = \frac{38}{64} \text{ or } 0.594$$

5.7.2 Yule's constant

On the other hand, Yule's Characteristic K is

$$K = 10^4(S_2/S_1^2 - 1/S_1) \tag{36}$$

where S_1 and S_2 are the first and second moments, respectively, of the word count distribution, that is the distribution of vocabulary items according to frequency of occurrence; $S_1 = \Sigma f_x X$; $S_2 = \Sigma f_x X^2$, where f_x is the number of words occurring X times. For large samples, and neglecting the factor 10^4 which Yule used only so as to avoid very small values of K, the Characteristic becomes

$$K^* = S_2/S_1^2 \tag{37}$$

which is a characteristic constant of the word count distribution epitomising the relation between vocabulary and word occurrence, that is frequency of use, in a given literary text (or author). It may be regarded as measuring the extent to which word occurrences are concentrated upon particular vocabulary items, and it shows that a particular style is characterised by a constant relation between uniformity and diversity in the number of repetitions of the items of vocabulary (Herdan, 1956).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

5.7.3 *The Repeat Rate*

Yule's Characteristic is thus a measure of *stylistic diversity*, and is mathematically expressed as the ratio of the second to the squared first moment of the word count, whereas Greenberg's constant is a measure of *linguistic diversity* among the members of a speech community and mathematically the sum of the squared probabilities of the various languages, which makes them look rather different from one another.

However, the close relation between the two becomes apparent if we use the interpretation given in a paper by Good (1953) of the Characteristic K as the repeat rate of words, i.e. the probability that two words chosen at random from a text will be the same dictionary word. That paper suggests methods of estimating, among other things, various general population parameters measuring heterogeneity. One such parameter for which Good uses the symbol $\hat{c}_{2,0}$ is an estimate of the probability that two words selected at random from the text under consideration will turn out to be the same word of the language. Hence, it tends to be larger the more repetitive is the author's vocabulary, or more roughly, the smaller is the author's vocabulary.

If the population probabilities of the distinct words are $p_1, p_2 \ldots$. then, as Good has shown

$$\hat{c}_{2,0} = \frac{S_2 - S_1}{S_1(S_1 - 1)} \tag{38}$$

represents an unbiased estimate of $p_1^2 + p_2^2 + \ldots$. as $S_1 \to \infty$. Since Yule's Characteristic K is equal to $10^4 c_{2,0}(1 - 1/S_1)$,

$$K = 10^4 \hat{c}_{2,0}\left(1 - \frac{1}{S_1}\right) = 10^4 \frac{S_2 - S_1}{S_1(S_1 - 1)} \cdot \frac{S_1 - 1}{S_1} = 10^4 \left\{\frac{S_2}{S_1^2} - \frac{1}{S_1}\right\},$$

it follows that it tends to $10^4(p_1^2 + p_2^2 + \ldots)$ as $S_1 \to \infty$.

Thus, Greenberg's measure of linguistic diversity A and Yule's Characteristic K are both of the nature of a repeat rate, and therefore of essentially the same mathematical structure.

5.7.4 *Illustration*

If the summation in the numerator and denominator of formula (37) refers not to f_x, i.e. groups of words of equal occurrence frequency, but to the individual words, the formula becomes

$$K^* = \Sigma X_i^2 / [\Sigma X_i]^2,$$

where X is the frequency of occurrence of a word, and different values of X imply different probabilities of occurrence. In Greenberg's formula for A, the corresponding probabilities of occurrence are the values of p_i for the different languages. Using the values of p_i from our hypothetical example and assuming a sample of the population of, say, 1000 inhabitants, the values of X result as

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\begin{aligned} 1/8 \times 1000 &= 125 \\ 3/8 \times 1000 &= 375 \\ 1/2 \times 1000 &= 500 \\ \hline & 1000 \end{aligned}$$

and the characteristic is calculated as

$$K = (125^2 + 375^2 + 500^2) / 1000^2 = 406.255 / 1{,}000{,}000 = 0.406$$

and

$$A = 1 - K = 0.594$$

which is the value of A calculated by formula (35) in accordance with the conclusion, reached on theoretical grounds, of the essential similarity in mathematical structure of Greenberg's coefficient of linguistic diversity *A* and Yule's stylo-statistical Characteristic *K*.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

VI

VOCABULARY CORRELATION

6.1 CONVENTIONAL MEASURES OF VOCABULARY RELATIONSHIP

Previous attempts at expressing quantitatively vocabulary connectivity between literary texts have used what is known in mathematical statistics as the correlation coefficient. According to whether only the number of vocabulary items or also their frequency of occurrence is taken into account, the method is that of the Bernoullian or the Pearsonian correlation (also called Product-Moment correlation).

The *Bernoullian correlation* measure is based upon the four word classes:

Words common to both texts,

Words appearing in one but not in the other text (2 classes),

Words appearing in neither text,

and thus depends for its magnitude not only upon the vocabulary both texts are connected by, but also on that which disconnects them from the rest of the literary material under consideration. It does not take the frequency of occurrence of particular words into account.

The *Product-Moment correlation*, on the other hand, takes frequency of occurrence into account when considering linguistic relationship between texts. For every vocabulary item the occurrence numbers in the two texts under consideration are recorded, and the magnitude of the correlation coefficient is the greater, the greater the number of vocabulary items with the same (or very nearly so) occurrence frequency in both texts relative to vocabulary items having different occurrence frequency. It is thus a measure not only of vocabulary connection, but also of frequency connection between the texts.

To illustrate the difference we may take the analysis of political vocabulary described in Vol. I, Section 4.1. Samples of the same length were taken from six texts of political material, the authors all belonging to the same historical period: Churchill, Halifax, Stresemann, Beneš, Stalin and Hitler. For each sample was determined, apart from its exact length in terms of number of words, the number of different vocabulary items it contained, and also the overlaps in vocabulary between any two authors. We then have, for each pair of authors, the four categories required for calculating the Bernoullian correlation coefficient, namely:

Words common to author A and B,

Words appearing in author A but not in B,

Words appearing in author B but not in A,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Words of the combined vocabulary of the 6 authors, not appearing in either A or B.

Since for each author the complete frequency distribution of the sample was recorded, i.e. the number of vocabulary items occurring once, twice, three times etc., it was also possible to calculate the Pearsonian correlation coefficient. The Product-Moment r is throughout higher than the Bernoullian r.

6.2 SHORTCOMINGS OF CONVENTIONAL METHODS FOR ESTABLISHING LINGUISTIC RELATIONSHIP

a. The Product-Moment correlation

Neither correlation coefficient is really very helpful in problems of stylostatistics, as we shall now show. We shall first deal with the shortcomings of the Product-Moment correlation coefficient, which Yule used in connection with the controversial question whether the *Imitatio* was a work by Thomas a Kempis or by Gerson. Working with three equal samples (sample size 2,454 words) from (a) the *Imitatio*, (b) other works by Thomas a Kempis, (c) works by Gerson, he determined the correlation coefficient between any two of these text samples. The correlation coefficient of the *Imitatio* and samples from Thomas a Kempis resulted as $r = 0.91$, that between the *Imitatio* and samples from Gerson as $r = 0.81$ and that between samples from Thomas a Kempis and samples from Gerson as $r = 0.84$, from which he concluded that the higher correlation between the *Imitatio* and samples from Thomas a Kempis was in support of Thomas a Kempis' authorship of the *Imitatio*. It can, however, not escape the observant reader that all three r are highly significant, and that it is only the difference between .81 and .84 on the one hand, and .91 on the other hand, which is used in this way in support of Thomas a Kempis' authorship of the *Imitatio*. This is, no doubt, not very satisfactory or logical, since it simply neglects the fact of the highly significant correlation between the other two groups of texts. This shows that speaking of literary relationship in terms of the magnitude of the Product-Moment r does not really do full justice to the problem of vocabulary connectivity.

Thus our conclusion in terms of the Product-Moment r about different or common authorship of literary texts is not according to whether r is, or is not, significant, but rests merely upon a comparison of whether r is greater or smaller in one bivariate distribution than in the other. Even unconnected texts may yield a significant r, and, conversely, even a non-significant r by itself does not necessarily lead to the conclusion of a diversity of authorship. This is hardly satisfactory.

In addition, there are other shortcomings of the method:

(1) It can only be applied to texts having the same length, or to samples of equal size from texts of different length, since it is only in equal pieces of text that we can expect a particular word to occur with more or less the same frequency.

(2) The comparison is restricted to two texts or samples at a time, without taking into account the larger whole to which they may belong.[1]

[1] This is not necessarily so.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(3) The computation is laborious, since it requires multiplying each bivariate frequency by the number of items belonging to both frequency categories simultaneously.

(4) The way in which frequency of occurrence enters the Product-Moment *r* has little to do with vocabulary connectivity (see below).

The method of ascertaining vocabulary connectivity which we shall propose to use is free from these disadvantages. As will be seen, it can be applied to texts (or samples) of unequal length, and to more than two such texts at the same time, since it necessarily takes into account the whole to which the two texts may belong; it provides the "yardstick" for lack of vocabulary connectivity beyond that due to mere random partitioning of vocabulary, and it is conspicuous for ease of calculation.

b. The Chi-square method (Bernoullian correlation)

To make us aware of the limitations of the Bernoullian *r* for determining vocabulary relationship, we shall use an investigation by C. D. Chrétien about "Word distributions in South-eastern Papua" (1956) which aims at determining linguistic relationship between populations of different geographical areas.

6.2.1 Illustration—the Melanesian languages of South Eastern Papua

Capell (1943) grouped these languages into 11 areas. From the total vocabulary he extracted 584 words as useful for comparative studies; these words fall into two classes: for 320 of them cognates can be found in Indonesian languages, and they are, therefore, taken to be of Indonesian origin. They occur in from 1 to 11 areas. The remaining 264 words do not have Indonesian cognates, but occur each in at least 2 of the 11 areas; they are, therefore, taken to be of Papuan (non-Indonesian) origin.

The following table shows the frequency of both classes of words according to whether they occur in 1, 2 .... 11 areas.

TABLE 26

Number of areas	Non-Indonesian		Indonesian	
	Number of words	Percentage	Number of words	Percentage
1	–	–	50	16
2	104	39	57	18
3	72	27	37	12
4	44	17	31	10
5	20	8	20	6
6	11	4	27	8
7	7	3	19	6
8	5	2	18	6
9	1	0	18	6
10	–	–	17	5
11	–	–	26	8
Total	264	(100)	320	(101)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Table 27 shows for each of the 11 areas what percentage of the 320 Indonesian words can be found (and not found), and similarly the percentages of the 264 non-Indonesian words present (and absent).

TABLE 27

Area	Non-Indonesian		Indonesian	
	%+	%—	%+	%—
1	14	86	45	55
2	23	77	72	28
3	45	55	53	47
4	28	72	34	66
5	25	75	43	57
6	33	67	38	62
7	47	53	39	61
8	48	52	52	48
9	21	79	42	58
10	30	70	45	55
11	14	86	23	77

6.2.2 *Establishing linguistic relationship by* χ^2

We are chiefly interested in the linguistic relationship between the areas. The number of possible pairs to be formed out of 11 areas is given by the formula for the number of combinations of 2 from 11

$$^{11}C_2 = \frac{n!}{r!\,(n-r)!} = \frac{11!}{2!\,9!} = \frac{10.11}{1.2} = 55$$

For each pair of areas, we now count the items of a given class of words (Indonesian or non-Indonesian) which occur

in both areas .... a
in one area only .... b
in the other area only.... c
in none .... d

This gives the 2 × 2 table on p. 100.

From a table of this kind we calculate a quantity called Chi-square (Sect. 18.5, Vol. I)

$$\chi^2 = N\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{39}$$

Since the formula for the (tetrachoric) correlation coefficient is

$$r = \frac{ad-bc}{\sqrt{[(a+b)(c+d)(a+c)(b+d)]}} \tag{40}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 28

		Area I		
		+	—	Total
Area II	+	a	b	$a+b$
	—	c	d	$c+d$
	Total	$a+c$	$b+d$	N

the relation between χ^2 and r is seen to be

$$\chi^2 = Nr^2$$
$$r = \sqrt{\chi^2/N} \tag{41}$$

If the .05 level of probability is adopted as a borderline between significance and non-significance, a Chi-square of 3.841 is required for significance for one degree of freedom, which is the appropriate number of degrees of freedom for a 2 × 2 table. Consequently, the minimum significant correlation coefficient is then found as

$$r = \sqrt{\chi^2/N} = \sqrt{3.841/264} = \pm 0.121$$

for the non-Indonesian component, and

$$r = \sqrt{3.841/320} = \pm 0.110$$

for the Indonesian component.

Table 29 gives some specimens of the values or r, a, b, c, d, $\%a$, $\%b+c$, $\%d$, for the non-Indonesian component. A similar table was obtained for the Indonesian data.

6.2.3 *Special vs. general components of linguistic relationships*

The method explained so far is the standard procedure for ascertaining the relationship of qualitative variables. But now attention must be drawn to a feature which is characteristic of its use in language statistics.

We know (Table 28) that numbers in the ++ and −− quadrant contribute to make the correlation positive, whereas numbers in the + − and − + quadrant make it negative, or at least weaken the positive correlation. But here, in Lexicon statistics, the numbers in the + + and those in the − − quadrant have a rather different import. The first imply the amount of agreement between the two areas in using the common source of non-Indonesian or Indonesian words. The second imply the extent to which they "agree to differ" from that common source, or the amount of rejection. The first may, therefore, be regarded as an index of the special relationship between two areas, and the second as epitomising their relation to the common or general source. We may, therefore, speak here of *special* and *general components* of correlation between two areas.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 29

Non-Indonesian Data (N = 264)

Pair	*r*	*a*	*b*	*c*	*d*	%*a*	%*b*+*c*	%*d*
1-2	.57	30	6	30	198	11	14	75
7-8	.53	94	30	32	108	36	23	41
10-11	.49	31	48	5	180	12	20	68
3-4	.32	53	67	22	122	20	34	46
5-9	.29	27	38	28	171	10	25	65
6-7	.28	59	29	65	111	22	36	42
4-6	.23	38	37	50	139	14	33	53
4-5	.21	29	46	36	153	11	31	58
3-10	.20	48	72	31	113	18	39	43
4-10	.16	31	44	48	141	12	35	53
5-10	.14	27	38	52	147	10	34	56
9-11	.12	12	43	24	185	5	25	70
3-5	*.11	36	84	29	115	14	43	44
3-6	*.11	47	73	41	103	18	43	39
3-9	*.11	31	89	24	120	12	43	45
4-9	*.09	20	55	35	154	8	34	58
6-8	*.08	47	41	79	97	18	45	37
6-7	*.07	15	73	21	155	6	36	59
5-11	*.05	11	54	25	174	4	30	66
9-10	*.05	19	36	60	149	7	36	56
...	.	.	.	.	.	.	.	.

* Non-significant values.

It is obvious that the determination of the linguistic relationship between any two areas does not do justice to the complexity of the problem. Relevant to the problem with which we are concerned would be not only the question of whether the vocabulary of a given area appears or does not appear in any other given area; more precisely, it does not only depend upon the four categories which we use for the calculation of the correlation coefficient for any two areas, but also upon how many words of the total vocabulary are peculiar to one given area, and how many are shared by it with 1,2, etc. of the other areas. This would lead to considering the vocabulary of each geographical area against the background of the ramifications of the total vocabulary among the different areas. However, such information is not available in the given experiment and we shall use in the following a substitute, however feeble, for the more correct procedure to be explained in the next Sections.

Chrétien's statistical method does not go beyond the listing of values as in Table 29, and analysing the results by individual comparison of the statistics *r*, *a*, *b*, *c*, *d*. The distinction of a general and special factor in linguistic correlations of this type suggests, however, trying whether more information might not be extracted from the data.

Having correlation coefficients for every possible pair of areas, it seems natural to

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

ask which area had, on the whole, the highest correlation with the other 10 areas. Under the restriction to be explained presently, that area could then be regarded as the most representative or typical of the lot. In general, the differences between areas in this respect might give a hint for their differentiation as regards language structure.

The totals of correlation coefficients (100 r) for each of the 11 areas are given in Table 30 (non-Indonesian and Indonesian).

TABLE 30

Sum of correlation coefficients of which a particular area, j, partakes

	$100 \sum_{t=1}^{11} r_{jt}$										
	Non-Indonesian										
Area No.	6	5	4	9	3	10	11	7	2	8	1
Total +	81	84	105	69	85	109	77	81	57	53	57
Total —	36	42	66	43	59	95	65	133	109	117	155
Algebraic Sum	45	42	39	26	26	14	11	42	52	64	98
	Indonesian										
Area No.	4	6	5	7	8	1	3	10	11	9	2
Totals	385	379	367	348	323	317	308	296	293	266	317

TABLE 31

Sum of values of d of which a particular area, j, partakes

$$\sum_{t=1}^{11} d_{jt}$$

	Non-Indonesian										
Area No.	11	1	9	5	2	4	10	6	3	7	8
Totals	595	577	553	536	525	510	494	496	398	373	371
	Indonesian										
Area No.	11	4	6	7	5	9	1	10	8	3	2
Totals	475	451	430	418	407	386	382	381	355	339	204

The restrictive condition mentioned above refers to the minus component of the positive correlation coefficient, that is the numbers denoted by d. They represent the words of the bulk not occurring in either area, and thus the part of the vocabulary with regard to which the two areas "agree to differ" with the bulk. Again we ask which area showed, on the whole, the greatest amount of such rejection. For this purpose we add the values of d similarly as was done for the values of r (Table 31).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

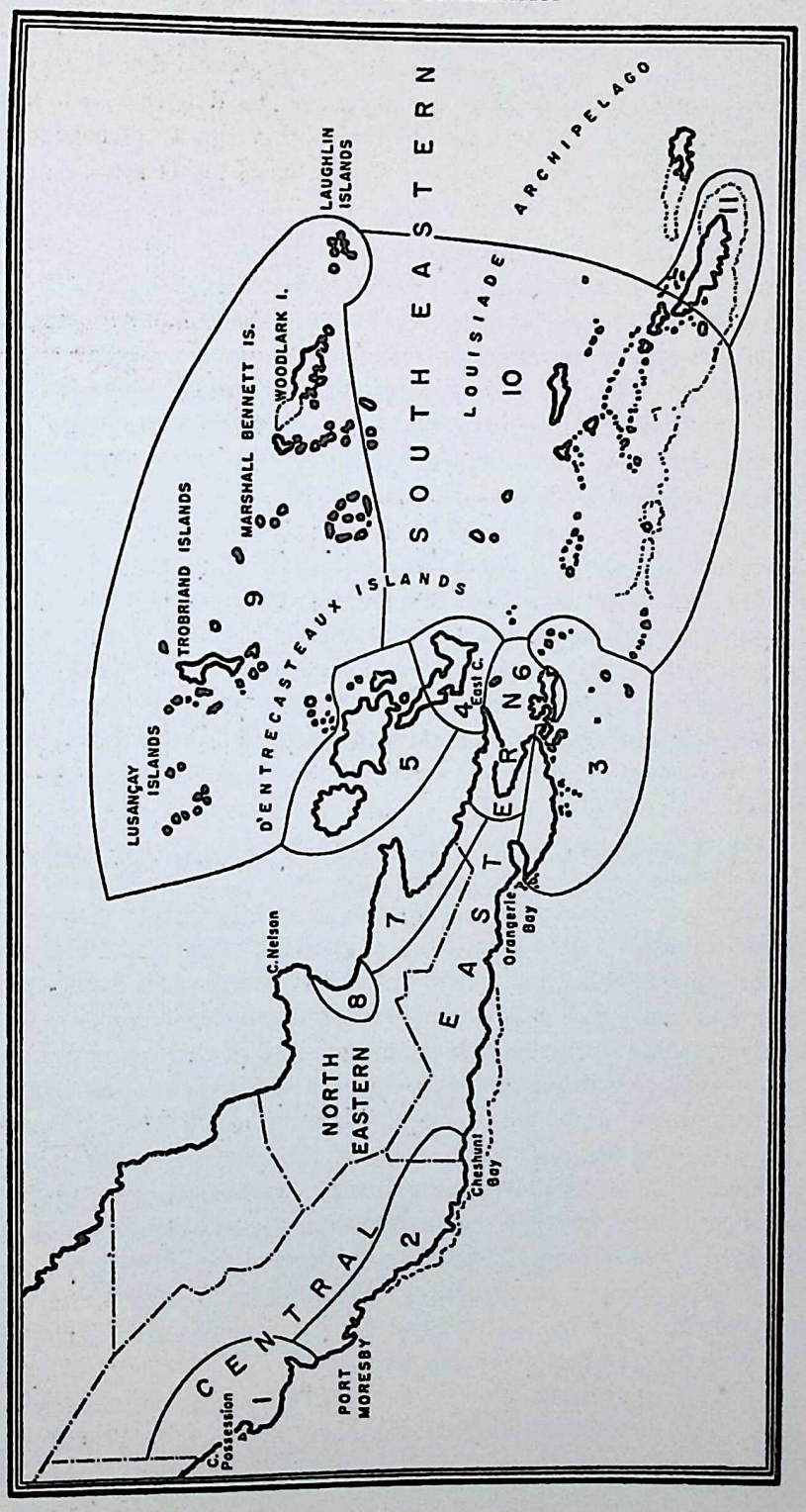


FIG. 9. South-Eastern Papua (after Chrétien)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Conclusions

Non-Indonesian. The areas arranged according to the magnitude of the sum of the correlation coefficients are

$$+ 6, 5, 4, 9, 3, 10, 11 \qquad - 7, 2, 8, 1$$

Looking at the map, Figure 9, we see that the difference in sums of correlation coefficients has its geographical counterpart. The areas whose sum of correlation *coefficients is negative lie west of a vertical line through Orangerie Bay and thus belong all to the mainland of Papua, the areas with a positive sum of the correlation coefficients lie east of that line,* forming the various groups of outlying islands, and the further east they are, the smaller is r. This would seem to indicate that there is greater linguistic homogeneity, or linguistic relationship as regards the non-Indonesian component in the region more remote from Indonesia.

Indonesian. Putting the series of r and d for the 11 groups side by side, we compute the ranking coefficient (Sect. 7.3 Vol. 1) which results as 0.5. The fairly high correlation between r and d leads to the conclusion that, on the whole, the correlation rests here upon rejection. The highest amounts of rejection are for the areas 11, 4, 6, 7, 5, 3 and smaller amounts for 1, 9, 10, 8, 2. *Apart from 11 and 10, the areas with high d are those in the centre of the map and those with smaller d are lying at the periphery.* We have thus here again a geographical counterpart for the overall difference of the cumulative statistic d for two groups of areas.

6.3. VOCABULARY CONNECTIVITY AND ITS USE FOR ESTABLISHING LINGUISTIC CORRELATION

The Bernoullian r has the important feature of taking into account that the two vocabularies examined are parts of a larger whole. It rests not only upon the vocabulary they have in common, but also upon that which is missing in both.

But strictly speaking, this is a rather summary way of characterising a part in relation to the whole to which it belongs, as we have seen in the previous illustration. A fuller characterisation would tell us also how many words one text has in common with 1,2, ... all the other parts of the whole, and this is what the statistics of vocabulary connection or vocabulary connectivity actually do for us. We shall first present this procedure as a method of descriptive statistics in which form it is used by R. Morgenthaler for describing the relationship between the different books of the Greek New Testament, and, in particular, of the different Pauline Epistles among themselves, and then show how it can be used as a correlation method in the strict sense of the term allowing scientific inference.

On the basis of his statistical dictionary of the Greek New Testament, Morgenthaler gives a table of the vocabulary connection of the 9 books of the New Testament: Matthew, Mark, Luke, John, Acts, Paul, Hebrews, Catholic Letters, Revelation, and

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

another of the Pauline Epistles in the following grouping: Romans, Corinthians (I & II), Galatians, Ephesians, Philippians, Colossians, Thessalonians (I & II). That is, he gives the number of words which are peculiar to any of these units and those which it shares with 1, 2, ... 8 other units. This enables him to discuss quantitatively—but without mathematical concepts and formulae—the relationship between these units by the token of vocabulary connection.

To make these data suitable for methods of mathematical statistics, we proceed as follows. Wishing to compare any two Epistles we must obtain for each the number of words peculiar to it, and those it shares with 1,2, ... the 8 other Epistles (we shall denote the groups according to the number of Epistles in each by the figures 1 to 9). The first category is given by Morgenthaler directly; the others must be obtained by summing over all Epistle combinations; he gives only the basic figures, that is the number of words which a given Epistle has in common with any other particular Epistle. Thus, for two-epistle words, we sum over 36 combinations, for three-epistle words over 84 combinations, for four-epistle words over 126 combinations, and so on. This gives for every Epistle 9 word categories, showing in the first and last how many words are peculiar to it and how many are common to all Epistles, respectively, and in the intermediate classes the vocabulary it shares with any 1,2, ... 7 of the other Pauline Epistles.

The pattern of vocabulary connectivity makes the tentacles of an octopus spring to one's mind, the first category, the words peculiar to a text, being the body, the other categories representing tentacles of different length.

The coincidence of there being, according to Morgenthaler's grouping, just 8 Epistles with which any particular Epistle shares its vocabulary makes the simile of the octopus still more suitable. The situation is thus one of a group of 9 vocabulary octopi, the body of each being represented by the one-epistle words, and the 8 tentacles of the 9 individuals being intertwined with one another. This sounds somewhat complex, admittedly, but, as will be seen, statistical analysis is quite capable of disentangling even such a seemingly hopeless chaos.

The 9 categories for each Pauline Epistle, according to Morgenthaler's grouping, obtained by summation of the basic data in Morgenthaler's Table 6c, are shown in table 32.

Having obtained the vocabulary distribution for each Epistle we can compare any two epistles by the Chi-square method. Since we take here into account the complete ramification of the total vocabulary among parts, the method can be regarded as the required generalisation of the tetrachoric correlation essential for a complete understanding of vocabulary connectivity. A significant Chi-square means that the distribution of vocabulary in the 9 categories of words according to their relation to the other Epistles, is such as would occur on pure chance less often than, say, 5 in 100 or 1 in 100 (according to the level of probability we prefer to use), or, as we say in modern statistics, that the Type I error is 5 in 100 or 1 in 100, meaning that using, say, the level 1 in 100, we shall be wrong once in 100 times with our assumption of a real

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 32

	One-Epistle words	Words shared with 1, 2,... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
Ro.	281	191	137	128	90	64	55	42	81	1069
Co. 1	246	156	132	117	75	71	55	40	81	973
Co. 2	177	136	96	106	60	57	43	40	81	796
Gal.	92	59	66	62	50	39	43	34	81	526
Eph.	83	78	75	48	43	39	44	38	81	529
Phil.	75	55	37	44	39	38	44	33	81	446
Col.	68	54	43	34	38	35	46	35	81	434
Th. 1 & 2	58	59	45	58	51	38	43	38	81	471
Pst.	335	134	98	76	52	51	40	36	81	903

difference. Conversely, a non-significant Chi-square implies that the Type I error attached to the conclusion of a real difference would be too great, and that we had better regard the two distributions as being only fluctuations due to random sampling of one and the same basic distribution.

It is important to realise the difference between the interpretation of Chi-square when obtained from the experimental design in the previous, and in the present, section. There we had *one* sample of words—the total vocabulary—distributed into four categories, and a significant Chi-square—and consequently a significant *r*—implied a significant *similarity* between the two geographical regions as regards vocabulary.

Here we have *two* vocabulary groups, each of which is distributed among nine categories, and a significant Chi-square—and *C* as a suitable measure of correlation—implies a significant *dissimilarity* between the two vocabulary patterns.

Before we apply the Chi-square test we must make sure, by a preliminary inspection of the data of Table 32 to what extent the test will give us the required answer. A striking feature of these data is that they appear to differ considerably from Epistle to Epistle in the first 6 columns, but not in the remaining 3. Pronounced differences between the entries of each column due to differences in the length of the literary text and possibly also due to differences in style, is just what we require for applying a Chi-square test in a sensible way, and this condition seems to be fulfilled for the first 6 columns. On the other hand, we find that the differences within the last 3 columns get smaller and smaller until in the 9th column the difference becomes zero, which shows that there are certain constraints at work which prevent text length, and possibly also style differences, from expressing themselves by corresponding differences in the number of vocabulary items. The nature of these constraints is quite obvious: column 9 contains the number of vocabulary items common to all Epistles and is thus by definition identical for all letters. The nature of the constraints thus lies in the greater number of Epistles sharing the vocabulary; these constraints will be somewhat less intense as we move away from column 9 towards the smaller Epistle combinations,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

but are still appreciable as far as column 7. As far as the contraints upon the number of vocabulary items are so powerful as to, more or less, obliterate the influence of text length and style, the data are not suitable for the Chi-square test. It is now a matter of somewhat arbitrary decision where to make the dividing line, and this necessarily must differ from experiment to experiment, but careful inspection of the table and the nature of the data would indicate that for the Pauline Epistles it seems advisable not to extend the Chi-square test beyond the first 6 columns.

This restriction of the Chi-square test to a specified number of columns, though less satisfactory from the standpoint of mathematical statistics, is yet fully justified in problems of the type we have before us. We know from Section 3.4 that most of the words in column 1 (one-epistle words) will be provided by the hapax legomena in the total, a great part of the numbers in column 2 by hapax legomena, dislegomena and some trislegomena, and in general that the number of words common to the small Epistle combinations will be mostly provided by words of rare occurrence. Against this the words common to the higher Epistle combinations, say from 7-9 in our case, will be provided by words with high and highest frequency of occurrence. Although it is a debatable point whether to use for establishing vocabulary correlation words with high or with low frequency, the latter alternative has more support from linguists, and rightly so. Words with high and very high frequency are those used by everybody and can thus not be usefully employed in establishing a connection between particular writers. This is where the weakness of the Product-Moment correlation coefficient lies when used for establishing linguistic relationship. Its value is more influenced by the words with high frequency. On the other hand, if two texts belonging to a larger whole share rarely occurring words, that is words which do not appreciably occur in the rest of the texts under consideration, this provides a useful argument for the vocabulary relationship between the two texts.

As is well known, the authorship of the Pastorals was perhaps the first problem to which statistical methods were applied on an extensive scale by Bible scholars. We are now in a position to contribute to the solution of that problem by the comparison of the pattern according to which the Pastorals share their vocabulary with the remaining Paulines, and the pattern with which any of the other Epistles share their vocabulary with the rest. Let us first compare the Pastorals with Romans (Table 33). Table 34a gives the Independence figures, that is the values expected on the assumption that the difference between the numbers for any one category of words was due only to the different text length of the two Epistles; they are calculated from the entries in Table 33 as the product of the corresponding marginal totals divided by the grand total. Table 34b contains the quantities calculated by the formula $\frac{(f_i - f_i')^2}{f_i'}$ where f_i stands for the observed frequencies of Table 33, and f_i' for the theoretical frequencies of Table 34a (Vol. I, Sects. 5.5.2, 18.5), and whose summation gives Chi-square. Summation of the entries in the first 6 columns gives a Chi-square of 33.50. It exceeds by far the Chi-square for the number of degrees of freedom (8 for the whole table)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 33

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
Romans	281	191	137	128	90	64	55	42	81	1069
Pastorals	335	134	98	76	52	51	40	36	81	903
Total	616	325	235	204	142	115	95	78	162	1972

TABLE 34a

Independence figures

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
Romans	334	176	127	111	77	62	52	42	88	
Pastorals	287	149	108	93	65	53	43	36	74	

TABLE 34b

Quantities whose summation gives Chi-square

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								
		1	2	3	4	5	6	7	8	
Romans	8.41	1.27	.78	2.60	2.19	.06	.17	0	.56	
Pastorals	1.96	1.51	.93	3.11	2.60	.08	.21	0	.66	

even at the .01 level of probability (20.09) and the difference in vocabulary connection of each of the two Epistles under consideration with the rest of the Paulines can therefore be regarded as statistically established. For our problem this means that the Pastorals are different in style from the Romans by the criterion of vocabulary connectivity (Table of χ^2 in Vol. I, p. 341).

Comparing now in the same way Romans with Corinthians I, we obtain for the first 6 columns a Chi-square of 3.22 which is far below the value required for significance. This can be interpreted as being due to there being no style difference in terms of vocabulary use between the two letters. Just one more illustration: the comparison between Romans and Galatians gives a Chi-square of 22.91 for the first 6 columns. In terms of significance and non-significance of Chi-square, we must conclude that the two Epistles are rather different in style, though much less so than Romans and the Pastorals. But it also shows that style difference and different authorship must not be taken as tantamount or identical concepts.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

In the previous section we derived from χ^2 the correlation coefficient r. This applies to χ^2 obtained from a 2×2 table. In general, for higher numbers of categories, χ^2 can be transformed into the Coefficient of Contingency by the formula (Vol. 1, p. 94)

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{42}$$

which is an index of correlationship in general, not like r one of linear correlation only.

For the Romans-Pastorals comparison we get $C = .09$, for Romans-Corinthians 1 $C = .12$, and for Romans-Galatians $C = .04$.

We must be careful not to confuse the meaning of C with nearness of relationship in the sense of similarity. On the contrary, C implies a systematic change in the figures corresponding to one another in the two series, whereas similarity of these figures, or rather of the proportions, implies the absence of a systematic change, and thus random fluctuations only. That C is more than twice as much for the Romans-Pastorals table than for the Romans-Galatians table is, therefore, to be interpreted as implying a systematic difference in the vocabulary connectivity pattern for the Romans on the one hand and for the Pastorals on the other.

6.4 VOCABULARY CONNECTIVITY VS. RANDOM PARTITIONING

The comparison of the data of Table 33 has so far been confined to any two Epistles, which enabled us to determine whether the vocabulary connectivity of each of the two Epistles in question with the rest of the Paulines were similar, or systematically different from one another. It is, however, quite obvious—and the smallness of the Coefficient of Contingency for our tables supports this—that even when the differences between two Epistles are such as to suggest a systematic difference in vocabulary connectivity with the rest of the Paulines, they still have much in common in this respect. For one thing, the trend of the figures in the direction from the one-epistle words to the eight-epistle words is similar for all Epistles. The question now arises where to draw the line between what we may regard as unavoidable similarity in the pattern and such dissimilarity as would result in a higher χ^2.

For the comparison of any two series we require, as we know, a table of Independence figures which serve as a standard for what would be the entries in the table on pure chance with the marginal figures and the total as given. Instead of comparing the vocabulary patterns of two particular Epistles, we may wish to compare the patterns of a given Epistle with what might be expected under the assumption of complete random partitioning of the total vocabulary of the Paulines among a number of parts of equal text length. This would enable us to determine the extent to which the pattern of vocabulary connectivity of the given Pauline Epistle deviates—and whether it does so significantly—from the pattern which might be expected on pure chance, that is in the absence of any particular connection between the Epistles.

Such a general pattern of vocabulary partitioning is obtained by means of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

random partitioning function which we encountered in Section 3.3, and whose use for establishing vocabulary correlation, or otherwise, will be discussed in greater detail in Chapter XVIII. Here we are only concerned with the pattern of vocabulary connectivity resulting from random partitioning. Having calculated by that method the random partitioning of the total vocabulary of the Paulines among 9 equal parts, we are now in a position to determine by means of the Chi-square test whether the vocabulary pattern of a given Epistle departs significantly from the random partitioning pattern of vocabulary connectivity.

Using the formula given in Chapter XVIII we obtain for the vocabulary pattern by random partitioning of the Pauline vocabulary among nine parts, the series:

TABLE 35

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
	132.5	105.6	86.8	78.4	70.0	61.6	58.8	68.0	174.0	835.7

Let us assume we wish to compare it with the vocabulary pattern for the Ephesians. For the purpose of such comparison, the figures of the random partitioning series must be changed proportionately so as to add up to the sum of the Ephesian vocabulary, that is we calculate the expected frequency of random partitioning in a sample of 529 words.

Since what we are interested in here is the difference between observation and expectation, the quantities whose summation gives χ^2 are calculated as the difference between the observed and the expected figure within any column, squared and divided by the expected figure.

TABLE 35a

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
Ephesians	83	78	75	48	43	39	44	38	81	529
Random partitioning series (adjusted)	83.9	66.9	54.9	49.6	44.3	39.0	37.2	43.0	110.1	528.9

$$\chi^2 = 18.82, \quad C = .133$$

which for 8 degrees of freedom is not significant at the .01 level of probability.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Comparing now the Pastoral series with what might be expected by random partitioning, we have the following table:

TABLE 35b

	One-Epistle words	Words shared with 1, 2,.... 8 Epistles								Total
		1	2	3	4	5	6	7	8	
Pastorals	335	134	98	76	52	51	4	36	81	903
Random partitioning series (adjusted)	143.2	114.2	93.8	84.7	75.6	66.5	63.5	73.4	187.9	902.8

Chi-square for the first and last column, that is for the one-epistle words and the words common to all Epistles is overwhelmingly high, viz.: 256.93 and 60.8 respectively, whereas for all the intermediate columns together it is 43.25. Summing these three values we obtain:

$$\chi^2 = 360.08, \quad C = .409$$

which leaves little doubt about the dissimilarity of the pattern of vocabulary connectivity for the Pastorals from what might be expected on random partitioning.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

SECOND PART

STATISTICAL LINGUISTICS

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

VII

THE DIFFERENT TYPES OF LANGUAGE STATISTICS

7.1 RANDOM SAMPLING IN LINGUISTICS

Random sampling as an operational concept was defined by Kendall and Babington Smith (1938) as a method of sampling for a specified characteristic such that the method is independent of the characteristic itself. They add: "There is no such thing as a random method of selection per se considered apart from the universe whose members are being selected .... Within the same universe, a method which is random in respect of one characteristic is not necessarily random in respect of another".

It follows that provided the method of sampling was independent of the characteristic for which we sample, it does not matter how regular or systematic the procedure is by which the individual samples are selected. This is what von Mises (1939) has called the "Prinzip des ausgeschlossenen Spielsystems" (the principle of the impossibility of a gambling system), which he regards as the most important criterion of a random series. If the universe is a random aggregate of events which may occur in two alternatives, say *A* and *B*, then no matter what elaborate system one were to devise for getting the better of chance, *provided only that the method did not imply a knowledge of the characteristics A and B themselves*, the resulting selection of items would again be a random series or random sample. Objections against v. Mises' principle as a criterion of randomness rest, as far as I have noticed, upon an incomplete knowledge of his definition, mostly forgetting the proviso.

My contention is that the linear sequences of linguistic forms in written texts or speech are random series with respect to certain quantitative characteristics,[1] and any sampling procedure, be it by disconnected linguistic units, or by continuous pieces of text, by pages or by chapters etc., will give a random sample of such a quantitative characteristic, provided only that the sampling method is in no way connected with the characteristic; that is, provided that it does not consist in a direct or indirect selection of categories of just the characteristic one is sampling for.

In Chapter IX it will be shown that what we call the population probabilities of phonemes is their functional burdening in the dictionary. Any method of sampling for phonemes with a view to obtaining their frequency distribution would, in the light of the above explanation, give a random sample, provided the method was neither

[1] See Sect. 20.3; The ergodicity of word masses.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

directly or indirectly such as to select certain phonemes, and neglect others. In a word, there must be no bias with regard to the type—and therefore the frequency—of phonemes which are comprised in the sample.

In Chapter X it will be shown that the frequency of occurrence of word length in speech and writing is a certain statistical function of the functional burdening of word length in the dictionary. It follows, in the light of our defining random sampling, that any method of scanning literary texts or speech for word length will give a random sample of the functional burdening, provided it is free from direct or indirect bias in selection.

In Chapter I, Sect. 1.5, we have spoken of the functional burdening of words with meaning, and of the relative frequencies of occurrence of words as reflecting their functional burdening. However, as Trubetzkoy said: "Man darf sich die Sache nicht zu leicht vorstellen."

Even if the universe of meanings were, as regards the functional burdening of words with them, a random variable, for which so far we can offer no direct proof, the selection of a certain text, necessarily dealing with a certain subject, is already something in the way of a biased selection, since the type-token relation expressed by K is to some extent dependent upon vocabulary, and selection of a text implies nolens volens selection of vocabulary. We must, therefore, expect some variation in K if the subjects are different, though this may be overcome by working with composite samples if carefully compiled.

For a particular author, or still better, a particular work, the sampling method for word frequency may be according to a random sampling technique, as regards the individual word occurrence or the lines or the page numbers etc., or it may be systematic, e.g. every tenth word, or the words on every 10th page, etc. Whatever method we choose, the sample of word frequencies will be a random sample of the population frequencies, which, as we believe, represent the functional burdening of the words with meaning, always provided that no direct or indirect selection of words as to frequency takes place. For instance, it was shown in Sect. 3.8, Vol. 1, that for two groups of vocabulary of Russian with very different frequencies of occurrence significantly different K's resulted. And in Sect. 5.4 the effect upon K of truncating the frequency distribution was demonstrated.

Apart from such selection, however, the particular semantic nature of the items comprised in the sample does not matter, that is, has no effect upon the constancy of K. This may be regarded as indirect evidence of the distribution of functional burdening of words with meaning as the statistical universe whose realisation are the samples of word occurrences of literary texts and of speech (see Sect. 1.5).

Incidentally, this hypothesis affords also an explanation of the fact that the stochastic model which Simon (1953) designed for a particular linguistic distribution seems to fit other, and rather different, distributions as well (Sect. 15.3.2). Simon explains this by the "weakness" of his assumptions. No matter how weak, they are in direct contradiction with the nature of some of those other distributions. A better explanation is

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

afforded by regarding the different types of distributions as random samples of one and the same universe, viz., that of the functional burdening of words with meaning. If the different distributions correspond to so many different ways of sampling, —without bias—of that universe, then we should expect them to obey the same law, and thus to be fitted by the same mathematical model (See Sect. 15.3.2).

7.2 VOCABULARY AND OCCURRENCE STATISTICS

Statistics have a two-fold purpose in linguistics: they might be intended to show how often a specified linguistic form (phoneme, phoneme combination, morpheme, word, etc.) occurs in "la parole", i.e. in speech and writing, or they might be intended to exhibit the use of such elements as phonemes, their combinations, morphemes in the vocabulary of the language, i.e. in "la langue", which leads to a classification of some sort or other of the material contained in the dictionary of the language. In other words: the "type" is the same in both vocabulary and occurrence statistics, but the "token" is in the first the frequency of "type" in the dictionary or in "la langue", and in the second the frequency of "type" in speech or in "la parole".

Attempts at statistics in language study often meet with the objection that since every member of the speech community uses the language in an individual way, language statistics could not go very far. The appropriateness of that objection depends very much upon the type of linguistic form which is made the unit of the statistical investigation. Roughly we may say that the larger the linguistic unit which is chosen for the investigation, the more will the frequencies depend upon individual style. To be more explicit: as we proceed from phonemes to phoneme combinations, to morphemes, to words, to sentences, the frequencies of occurrence of particular units appear more and more to be conditioned by individual style. Or better perhaps: the more a linguistic form is connected with meaning, the greater the influence of the individual style. As an illustration may serve samples from two German authors (Tables 36, 37).

Number of phonemes, their consonant and vowel parts, word length, number and place of consonant combinations, all may be connected with style differences. The scholarly style (Bühler) is characterised by long words, rich in consonant combinations (through the use of foreign words), the popular style (Dirr) prefers short and simple words.

The intellectual, scholarly style and primitive, descriptive style are the two poles, or extremes, between which the various types of style find their place. Every text has its own style. To find a completely representative text in this respect, free from the influence of a particular style, will hardly ever be possible. The way out of the difficulty is to work with composite samples.

The linguistic characteristics mentioned above—and others—are therefore style-conditioned.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 36

Illustration (after Trubetzkoy)

Word length in terms of no. of syllables and nos. of vowels and consonants

No. of syllables	K. Bühler (Scholarly style)		A. Dirr (Primitive style, Translation of Fairy Tales)	
	No. of words		No. of words	
	Absolute	Relative (in %)	Absolute	Relative (in %)
1	95	47.5	134	67
2	57	28.5	56	28
3	27	13.5	10	5
4	7	3.5	—	—
5	6	3.0	—	—
6	6	3.0	—	—
7	1	0.5	—	—
8	—	0.0	—	—
9	1	0.5	—	—
	200	100.0	200	100

Total of syllables	400	276
Average length of word	400/200=2	276/200=1.38

Since in German a syllable must contain at least one vowel, the vowel numbers correspond to the syllable numbers, and we have

Vowels	400	276

The consonants were counted and their numbers are:

Consonants	636	429
Total	1036	805
Consonants/vowels	63%	64%
Consonants/total	61%	61%
Vowels/total	39%	39%

We must, however, beware of generalising from this instance and conclude that word length, be it in terms of syllables or of letter or phoneme number was really so very much dependent upon style. The two text samples under consideration, one from a scholarly treatise, and the other from a collection of fairy tales, were selected with a view to obtaining a maximum of difference in word length and can, therefore, not be regarded as representative of the word length distribution in the *language*, or even as one which would result from a very large sample. In fact, it can be regarded as established that word length in terms of syllable number has a fairly stable distribution if the samples are not deliberately chosen from texts whose subject matter necessitates wide differences in this respect (see Sect. 10.3, 10.5 Vol. 1).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 37

(after Trubetzkoy)

	Distribution of vowel phonemes in the same samples from							
	K. Bühler No. of Occurrences				A. Dirr No. of Occurrences			
a			15		18			
ə		37				40		
			22		22			
	57						63	
u			7		9			
o		20	10		10	23		
au			3		4			
i			17		16.5			
e, ä		39	18		11	35		
ei	43		4		7.5			
ü			3		1		37	
ö		4	0.5		1	2		
öü			0.5		0			
			100		100			

The relative frequencies of particular phonemes in the two samples are very similar and, if the samples were larger, may be shown as differing by not more than could be accounted for by fluctuations of sampling.

It has been abundantly established that phoneme distributions from even widely different texts, as far as subject matter is concerned, can be regarded as random samples from one basic distribution. Thus phoneme frequency does not appear to be style-conditioned, and we say, therefore, that phoneme frequency is to a very large extent language-conditioned.

7.3 THE DISTRIBUTION OF PHONEMES WITH REGARD TO WORD STRUCTURE—OBSERVED VS. POSSIBLE FREQUENCY OF PHONEMES

A characteristic feature of true language statistics, capable of providing linguistically relevant information, is that they do not stop at collecting figures about language units, but always aim at the comparison with a standard. This enables us to ascertain the presence and the degree of selective variation, as distinct from all possible, or the most probable, occurrences of the linguistic event in question. The type of standard differs according to the material. Thus, it is not enough to have ascertained the relative frequencies of the different phonemes in a language, but we ought to make such information the basis for linguistically relevant comparisons; for instance, the comparison of the frequency of a given phoneme in phonetically relevant position with

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

its global frequency, or the number of different vocabulary items in which the phoneme has distinctive function with the total number of vocabulary items containing that phoneme. In short, we could utilise the global frequency as a basis for deriving information about the functional burdening of phonemes. Here the standard is provided by the *global* frequency of a given phoneme. In other cases, the standard may be the number of *possible* occurrences of a phoneme in a certain position of the word, against which to judge the actual or observed number of words having the phoneme in that position. A third type of standard is the *probable* range of variability of a certain language feature, that is its variation according to a law of chance, with which to compare the actual or observed variation of that language feature; according to the outcome of the comparison, we may conclude

(a) that the feature in question varied no more than could be accounted for by chance only, and thus that it is purely random;

(b) that there is an assignable cause at work which made it vary in excess of what was possible by chance only; and

(c) that there existed a linkage between certain linguistic features which significantly reduced the variation due to chance.

Instead of setting up hypotheses about the use of phonemes in general (about differential phoneme frequencies—Trnka, Zipf), which Trubetzkoy rightly regarded as "aussichtslos", it is more profitable to compare the observed with the combinatorially possible figure minus the number of inadmissible combinations.

In obtaining these figures, all the rules must be taken into account about cancellation of the phonological opposition in certain circumstances, i.e. in specified positions of the word or syllable. This means that it is the laws of sound clustering peculiar to the language which here provide the standard with which to compare the observed frequencies.

For instance, suppose that in a language the "merkmallose" member of the opposition concerned may occur at the *beginning of a word* and *within a syllable*, and the "merkmaltragende" member only at the beginning of all syllables except the first. If the average number of syllables per word is a, then we may expect that the frequency of the "merkmallose" member be related to that of the "merkmaltragende" as $(a + 1)$ to $(a - 1)$ (Trubetzkoy, 1939).

In the Chechen (Tschetschenisch) language the opposition of geminated and single consonants is in agreement with our assumption. We should, therefore, expect, since word length is 1.9 syllables approximately, that the frequencies of geminated and single consonants be in the ratio 0.9 : 2.9 or 1 : 3 approximately. Actually we have

tt : *t*	12 : 90	or	4 : 30
qq : *q*	6 : 45	or	4 : 30
cc : *c*	25 : 59	or	13 : 30
ll : *l*	16 : 32	or	15 : 30

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Geminated *cc*, *ll* are thus more frequently used, and *tt*, *qq* less often, than expected.

Another example from the Chechen language.

The "merkmaltragende" member of a certain opposition can be used only in an initial position of words; if the average syllable number per word is *a*, and the number of syllables in a text is *b*, its expected frequency is then b/a. The corresponding "merkmallose" member may occur in the initial position of a *syllable*, and also in the final position of a *word*, and thus $b + b/a$ times. The expected (better combinatorially possible) frequencies are thus related like $b/a : (b + b/a)$ or like $1 : (a + 1)$, and since $a = 1.9$, like 1 : 3 approximately. Actually, we have

$t' : t$	33 : 90	(11 : 30)
$k' : k$	38 : 47	(24 : 30)
$q' : q$	21 : 45	(14 : 30)
$c' : c$	17 : 97	(5 : 30)
$č' : č$	5 : 59	(2.5: 30)

For the total we have the relation as 114:365 or 1:2.97 which is in good conformity with the expectation 1 : 2.9 or 1 : 3 approx.

Only in relation to the possible frequency do the observed frequencies acquire a value for the linguist, because it is this relation which shows whether a phoneme is much or little used in the language. The ratio $\frac{\text{observed}}{\text{theoretical}}$ frequencies is a measure of the *functional burdening* of a phoneme or the phonological opposition.

All this shows the need for replacing the global phoneme statistics by a finer and more specific one, whose object is no longer the phoneme but the phonological opposition.

7:4 PHONOLOGICAL STATISTICS OF VOCABULARY—FUNCTIONAL BURDENING AND UTILISATION OF PHONEMES

Statistical investigation of texts is by itself not sufficient for getting an idea of the relative utilisation (functional burdening) of the different phonemes. It must be supplemented by phonological statistics of the lexicon. (But the reverse is also true; see below).

A start in this direction was made by G. V. Mathesius and B. Trnka. What may be achieved in this way is the *phonological characterisation* of languages. E.g. In German 18 consonants (see Table 38) may be in initial position in words and 14 in the final position, whereas the accented (betoned) vowel phonemes (10—if *ä* and *e* are not distinguished) only occur in both initial and final positions; the following combinations are not admitted: $j + i$, $au + r$, $au + ŋ$, $eü + r$, $eü + ŋ$, $ai + r$, $ai + ŋ$, $o + ŋ$, $ö + ŋ$.

Thus, theoretically, there are possible in German $(18 \times 10) - 1 = 179$ words of the type consonant + vowel, and $(14 \times 10) - 8 = 132$ of the type vowel + consonant.

Actually, we find in German 57 (one-syllable) words of the first type, and 37 of the second type.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 38

2-Phoneme Words in German	
Consonants	Initial (18) *b p m d t n k g c z s f v p h r l j* Final (14) *p m t n k y x c s p f s r l*
Vowels (10)	which may appear either initially or finally, but certain such combinations (9) are not admissible.
Theoretically possible	$(18 \times 10) - 1 = 179$ words of the type $C + V$ and
2-phoneme words:	$(14 \times 10) - 8 = 132$ words of the type $V + C$
Observed	$C + V = 57$ $V + C = 37$
Observed/Theoretical	$C + V = 31.8\%$ $V + C = 28\%$

2-Phoneme Words in French		
Theoretically possible number	(15×13)	$= 195$ words of the type $C + V$
	$(12 \times 18) - (4 \times 6)$	$= 192$ words of the type $V + C$
Observed	$C + V$	$= 142$
	$V + C$	$= 50$
Observed/Theoretical	$C + V$	$= 73\%$
	$V + C$	$= 26\%$

In French, 15 consonants are allowed to be in the initial, and 18 in the final position, whereas 12 vowel phonemes may occur in closed syllables and 13 in open ones. 24 combinations are not admitted. Theoretically possible are thus $15 \times 13 = 195$ words of the type consonant + vowel, and $(12 \times 18) - (4 \times 6) = 192$ of the type vowel + consonant.

Actually we find 142 words of the first type and 50 of the second.

This gives the following comparison of the *relative* i.e. observed/possible frequencies:

	cons. + vow.	vow. + cons.
German	31.8%	28%
French	73%	26%

All such peculiarities may be expressed in figures. Very important is the numerical estimate of the different functional burdening of the phonological oppositions, and the "average burdening" of phonemes in general. It then appears that languages may, in this respect, be "economical" or "wasteful". In the former type, words distinguished by a single phoneme are very numerous, and the percentage of realisation of the theoretically possible phoneme combinations is high. In the latter, there is a tendency to distinguish words through more than one phoneme and only a part of the possible combinations is realised.

With all this in mind, phonological statistics of texts acquires a new meaning. The

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

frequency figures have apparently *a double relativity*: the statistical investigation of *vocabulary* reveals the percentage of realised phoneme combinations in the language; i.e. which combinations have a definite semantic import, and with it the "functional burdening"; the *statistical investigation* of *word occurrences* in texts reveals the differential *utilisation* of the realised phoneme combinations. It may happen in a language that word types with a high percentage of realisation of the theoretically possible phoneme combinations have a lower frequency of occurrence (in "la parole") than words with a small percentage of realisation.

In Chapter VIII the relation between functional burdening and utilisation in speech is subjected to a closer scrutiny.

Lexico-statistics meet with similar difficulties as phonological statistics, but Trubetzkoy recommends the application of statistical methods also to vocabulary problems.

7.5 A GENERAL CALCULUS OF PHONEME DISTRIBUTION IN THE DICTIONARY

In a more general way, the phonemic distribution can be described by using the mathematical theory of sets, also known as 'relation theory'. As will be seen, we may regard the distribution of phonemes as a model in linguistic sets for the theoretical constructs of the mathematics of sets (Harary and Paper, 1957).

Any occurring sequence of phonemes belonging to the Set P of phonemes will be called an "ordered pair". P is the number of phonemes in the language plus an element # for "space" or plus "junction". The implied relation R between the members of an ordered pair may be described as "is immediately followed by".

We shall speak of the two positions in any ordered pair as the "predecessor", or α-position, and the "successor" or β-position. By αx we denote the sub-set of members of P that occur in the first position (α) where a specified phoneme x holds the second position; and similarly we denote by βx the sub-set of members of P that occur in the second position (β) when a specified phoneme x holds the first.

There are six possible relations: 1. Completeness, 2. Symmetry, 3. Antisymmetry, 4. Reflexivity, 5. Irreflexivity, 6. Transitivity.

Of these we shall deal in detail only with 1, 2 and 3, and illustrate the method by samples from Japanese.

1. *Completeness.* By α-completeness of a phoneme x we understand the sub-set of P which occurs in all ordered pairs of which x is the second member, e.g. in Japanese the phoneme *s* can be preceded by *a*, *o*, *u*, *e*, *i*, #, *n*, *š*, that is by seven phonemes plus juncture. Similarly, the β-completeness of a phoneme means the sub-set of P which occurs in all ordered pairs of which x is the first member. In Japanese the phoneme *s* can be followed by *a*, *o*, *u*, *i*, #, *č*, *t*, *k*, *c*, *š*, that is by nine phonemes plus juncture.

By "total binary interaction" or t-field of a phoneme x we understand the sub-set of P which belongs to either the α- or β-field of x or to both.

To find a numerical expression for completeness of a phoneme we divide the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

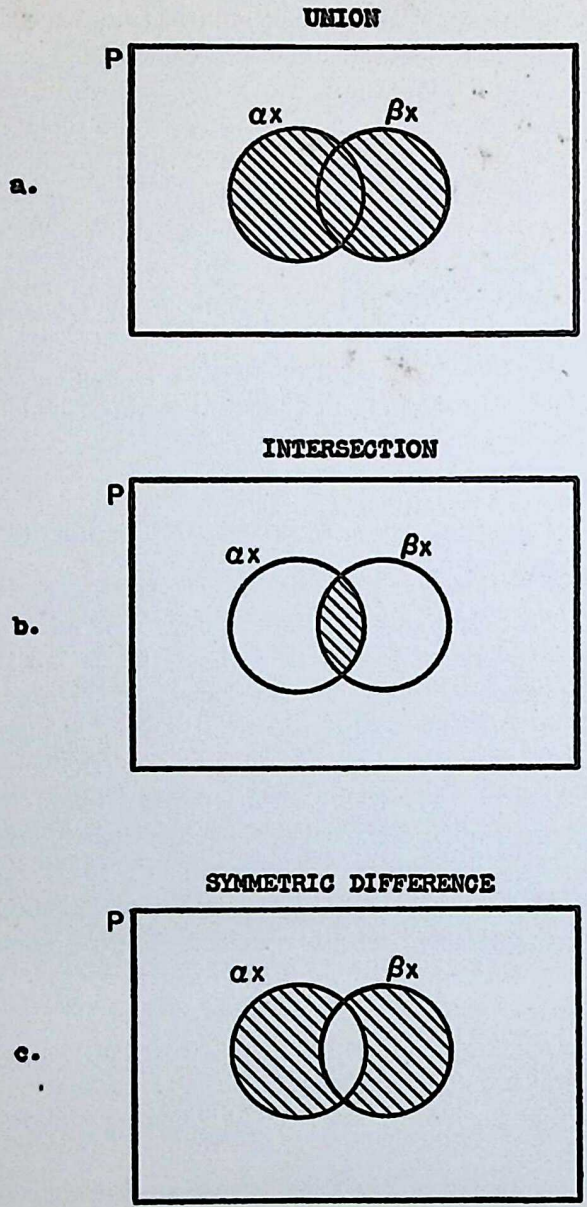


FIG. 10. (after Harary & Paper)

number in the sub-set by the number of phonemes plus junction $P = 29$. Writing for completeness of the relation R, $K(R)$, we have

$$K(R)\alpha = \tfrac{8}{29}$$
$$K(R)\beta = \tfrac{10}{29}$$

For the completeness in the combined field t, the sum of α and β must be diminished by

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

½ the ordered pairs in the intersection (see fig. 10) which are counted twice. There are 12 such pairs and we have, therefore,

$$K(R)_t = \tfrac{8}{29} + \tfrac{10}{29} - \tfrac{6}{29} = \tfrac{12}{29}$$

Instead of dividing by P, we might form the ratio of the sub-sets α and β to the total interaction, that is to the number in the t-field. This is called the internal completeness. For the phoneme s in Japanese we get this internal completeness in the α-field as $\frac{8}{12}$ and in the β-field as $\frac{10}{12}$.

The degree of completeness for the whole phoneme inventory of the language is the ratio of the sum of numbers of ordered pairs for all phonemes to the square of the number of phonemes in the language, P^2, minus 1. P^2 represents the possible number of combinations of elements, but since the sequence # # cannot possibly occur, we must subtract unity. We thus write

$$K(R) = \frac{\sum_1^{20} K(R)_t}{P^2 - 1} \tag{1}$$

For conservative Japanese we get $\frac{330}{840} = .39$ and for innovating Japanese $\frac{366}{960} = .38$.

2. *Symmetry.* The phoneme x is called symmetric with respect to a phoneme y if both sequences /xy/ and /yx/ occur. The phoneme x is a "symmetric phoneme" if it interacts *symmetrically* with all phonemes with which it interacts. The degree of symmetry is again expressed as the ratio of the number of phonemes with which a phoneme x is symmetric to all phonemes in P; the degree of "internal symmetry" by the ratio of the number of phonemes with which it interacts symmetrically to the number of phonemes with which it interacts (Fig. 10).

3. *Antisymmetry.* A phoneme x is said to be unsymmetric to a phoneme y if both /xy/ and /yx/ do not occur. Such ordered pairs are represented by the segments of α and β circles outside the intersection. The degree of antisymmetry of a phoneme is $^1/_2$ of the sum of ordered pairs outside the intersection—because each ordered pair xy is counted twice, once for the predecessor phoneme and once for the successor phoneme—divided by the number of ordered pairs in the t-field.

The degree of antisymmetry for the whole phoneme inventory is then given as

$$S'(R') = \frac{\frac{1}{2}\sum_1^{20} \text{No. of ordered pairs outside the intersection}}{\text{No. of ordered pairs for all phonemes}} \tag{2}$$

For Japanese we have:

		Conservative	*Innovating*
Symmetry	$S(R)$	$\frac{243}{330} = .73$	$\frac{276}{366} = .75$
Antisymmetry	$S'(R)$	$\frac{88}{330} = .27$	$\frac{90}{366} = .25$

7.6 THE FREQUENCY DISTRIBUTION OF PHONOLOGICAL OPPOSITIONS

A. *Classification of oppositions* according to their relation to the whole sphere of oppositions.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(1) *One- and multi-dimensional*, according to whether the "basis of comparison" i.e. the total of properties they have in common, is peculiar to them or not. In German $t - d$ is one-dimensional, but $d - b$ is multi-dimensional, since the common feature is also found in g. In the German consonant system, 13 oppositions are one- and 177 multi-dimensional.

(2) *Proportional and isolated*, according to whether the relation between the members is identical with that between other phonemes in the same language. In German, $p - b$ is proportional to $t - d$ and $k - g$. The German "Bühnensprache" has 20 consonant phonemes and thus $20!/18!2! = 190$ possible consonant oppositions. Of these 40 oppositions are proportional and 150 isolated.

Combining (1) and (2), the bi-variate distribution is as follows (Trubetzkoy):

		proportional	isolated
Dimension	one	11 (6%)	2 (1%)
	more	29 (15%)	148 (78%)

In the one-dimensional group, the proportional ones are more numerous, and in the multi-dimensional the isolated.

The absolute figures and percentages differ from language to language, but the general trend is the same: the isolated multi-dimensional form the biggest group, the isolated one-dimensional the smallest; between them are the proportional oppositions, among which the multi-dimensional are always more numerous.

B. For characterising a given system, the *proportion of phonemes* taking part in the various types of opposition is more important than the percentage of those types.

Thus in German only one phoneme, *h*, occurs exclusively in isolated, one-dimensional opposition, and three, (*sch*, *r*, *l*) in multi-dimensional isolated. All the rest, 16 out of 20 (80%) take part in proportional opposition. For Russian, the figure is 88%, for Burmese 97% (Trubetzkoy).

C. Even more characteristic is the relation of the number of one-dimensional proportional oppositions to the number of phonemes taking part in these oppositions, as expressed by the ratio: opposition number/phoneme number. That ratio is:

for German ..... $\frac{11}{16} = .69$
for Russian ..... $\frac{27}{30} = .90$
for Burmese ..... $\frac{79}{60} = 1.32$

Observe: under (*B*) ratios are formed of the number of *phonemes* taking part in specified types of *oppositions*; under (*C*) ratios are formed of the number of *oppositions* over that of *phonemes*. This is a cross between (*A*) and (*B*). Under (*A*) ratios are formed of the number of oppositions of different type.

"*Correlation*": if the opposition between two phonemes is uni-dimensional, proportional, privative and can be cancelled (aufhebbar), they must be very closely or intimately related; if the opposition is multi-dimensional, isolated, and thus cannot be cancelled (unaufhebbar), they are maximally remote from one another. These represent the extremes of high and low correlation.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

THE RELATION BETWEEN THE FUNCTIONAL BURDENING OF PHONEMES AND THEIR FREQUENCY OF OCCURRENCE

The distinction between the two types of relative frequency of phonemes, viz. as the ratio of occurrence numbers to text length (in terms of total phoneme number) and as the ratio of their incidence in the dictionary to the total of phonemes there, is not new. It was probably stressed for the first time by Trubetzkoy in *Grundzüge der Phonologie* where he speaks of the "doppelte Relativität" in phoneme counts. However, the relation between the two distributions of relative frequencies has so far not been reliably established. Consequently, what we read about it is mere guesswork, based upon subjective impression—just the thing we want to avoid in quantitative linguistics as an empirical science. As a rule, the two distributions are regarded as basically different; even so unsurpassed an expert in matters of phonology as Trubetzkoy seems to have taken this for granted, since otherwise he would not have used the term "double relativity" (Trubetzkoy, 1939).

The purpose of this investigation is to study the relation in question using the methods of quantitative linguistics which are now available to us.

8.1 THE PHONEME DISTRIBUTION OF ENGLISH ACCORDING TO THE DICTIONARY

A very detailed investigation of English mono- and dissyllabic words in the Oxford Dictionary was carried out by Trnka in his book *A Phonological Analysis of Present-Day Standard English* (1935). That it is limited to mono- and dissyllables does not materially detract from the generality of its results, because words of more syllables (many of which are of foreign origin) are relatively infrequent in both the spoken and the written language: according to Dewey (1923), their frequency of occurrence does not exceed 8%. We shall return to this point when the question arises of what this means in terms of dictionary incidence.

On the basis of Trnka's work a detailed quantitative analysis of phonemes in English was carried out by Krámský, who grouped the English phonemes of the dictionary material according to the manner of articulation and according to the place of articulation or the organ of production. According to the manner of articulation he divides the 23 English consonant phonemes into plosives and affricates: / b d g p t k tʃ dʒ /, fricatives: / θ ð f v s z ʃ ʒ and h /, and nasals and liquids:

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

/ l m n ŋ r w j /. According to the place of production they are divided into labials: / p b m w f v /, dentals and alveolars: / θ ð t d n l s z ʃ ʒ r /, palatals: /j/, velars: /k g ŋ/ and glottal: /h/. The results may be summarised as in Table 39.

TABLE 39

Dictionary distribution of English consonant phonemes

		Monosyllables	Dissyllables	Mono- and dissyllables
Manner of articulation	Plosives & affricates	3122 (40.8%)	2579 (41.4%)	5701 (41.1%)
	Fricatives	1765 (23.1%)	1234 (19.9%)	2999 (21.6%)
	Nasals & liquids	2757 (36.1%)	2405 (38.7%)	5162 (37.3%)
	Total	7644 (100.0%)	6218 (100.0%)	13862 (100.0%)
Place of articulation	Labials	1849 (24.2%)	1528 (24.6%)	3377 (24.4%)
	Dentals	4525 (59.2%)	3771 (60.6%)	8296 (59.8%)
	Palatals	106 (1.4%)	101 (1.7%)	207 (1.5%)
	Velars	1164 (15.2%)	818 (13.1%)	1982 (14.3%)
	Total	7644 (100.0%)	6218 (100.0%)	13862 (100.0%)

We note that there is a considerable similarity in the phoneme distribution, both according to the way and according to the organ of production, for monosyllables and dissyllables. Incidentally, this explains why the quantitative text characteristic of word length in terms of syllable number which, as Trubetzkoy has already pointed out, is partly dependent on individual style, does not interfere with the far-reaching similarity of the phoneme distributions in texts from widely different writers (Sects. 7.2, 7.3).

The similarity of the two phoneme distributions in mono- and dissyllables, especially according to the place of production, is such that we may regard them as random samples from one statistical population. We have therefore added to Krámský's table the distribution in both mono- and dissyllables taken together, which we regard as a series of population frequencies or probabilities.

8.2 THE UTILISATION OF ENGLISH PHONEMES IN SPEECH

Corresponding figures for phoneme frequencies according to place of production in running texts are available from quite a number of investigations. We take as a representative sample the figures from an investigation by Fowler (1957). Four samples were examined, all chosen from modern English. Each sample was transcribed into a phonemic alphabet of 24 consonants and 6 vowels. Three of the samples

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

consisted of 5000 consecutive phonemes each; the fourth, a story for children, is complete in 501 phonemes. The samples of 5000 were each divided into 5 samples of 1000 each, and the mean, standard deviation and coefficient of variation of the sampling distribution of means, v_m, were calculated for both the total and the partial distributions. The results show that the type-token relation of phonemes is not significantly disturbed by individual differences in style (author or subject).

The advantage in using Fowler's work is that each text sample of 5000 phonemes is divided into 5 samples of 1000 phonemes each, which enables us to test the stability of phoneme distribution within and between samples. Grouping the consonant phonemes in the running text samples according to place of articulation in the same way as those of the dictionary material have been grouped, we get the distributions shown in Table 40. Dewey's data, shown in the last column of Table 40, give a very similar distribution.

TABLE 40

English consonant phonemes—running text count

	Concepts of the Calculus by C. B. Boyer			*Fielding & Sterne* by G. Greene	Average of two preceding columns	Dewey
	2nd 1000 phonemes	4th 1000 phonemes	1st 5000 phonemes	1st 5000 phonemes		
Labials	144 (23.2%)	135 (22.2%)	694 (22.7%)	750 (24.2%)	722 (23.5%)	21.3%
Dentals	407 (65.5%)	403 (66.4%)	1996 (65.4%)	2002 (64.5%)	1999 (64.9%)	67.3%
Palatals	6 (1.0%)	6 (1.0%)	30 (1.0%)	31 (1.0%)	30.5 (1.0%)	1.0%
Velars	64 (10.3%)	63 (10.4%)	334 (10.9%)	321 (10.3%)	327.5 (10.6%)	10.4%
Total	621 (100%)	607 (100%)	3054 (100%)	3104 (100%)	3079 (100%)	100%

The far-reaching similarity between the distribution of consonant phonemes in the 1000 and 5000 phoneme samples from the same text, between the 5000 phoneme samples from different texts and between these and the very great sample from Dewey (22,477 consonant phonemes) is not surprising. The implied stability of phoneme occurrence in running texts has been established before (Herdan, 1956). What is remarkable is the close similarity between the phoneme distribution for *dictionary material* and for *running texts*, which can be seen by comparing Tables 39 and 40. However, it would appear that when used in speech, dentals are in excess, and velars in defect, of the corresponding dictionary frequencies. This could mean that there are two factors involved in the stability of relative frequency of phonemes in use: the pattern provided by the functional burdening of phonemes (in the dictionary) and, though to a much lesser extent, the need for divergence from it.

To test this assumption, we shall first try to derive the frequencies of phonemes in speech by a stochastic process from the pattern of functional burdening in the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

dictionary, and see what conclusions can be drawn from such numerical agreement as we obtain between the two series—text occurrence and dictionary incidence, or frequency of use of phonemes in speech and their functional burdening in "la langue".

Incidentally, the similarity between the two distributions admits the conclusion that the words of three or more syllables in the dictionary will not exceed 8%, the value for the frequency of occurrence of such words in the text, which enables us to assess the error we are making by working with mono- and dissyllables only.

8.3 THE DERIVATION OF THE OCCURRENCE DISTRIBUTION FROM THAT IN THE DICTIONARY

The stability of the distribution of frequency of use, its independence of the type of text, points to a general function according to which phonemes are produced in speech. Insofar as the phoneme distribution from texts is similar to that from the dictionary, it might well be considered to be a random sample of the latter, and a law of chance might present such a general function. We shall therefore use the Poisson law of rare events, which is most likely to fit the data in question if they are governed at all by chance, and calculate the numbers of phonemes belonging to the different categories according to place of production (Yule, 1944). If the calculated numbers sensibly agree with those actually present in the sample, and if the proportions are also like those observed, we would conclude that the frequencies of use of phonemes may be regarded as random samples of their functional burdening, and the distributions of the former as random samples of the distribution of the latter.

For each of the four phoneme groups the number, n, of consonant phonemes to be expected in a text of specified length, N, is calculated according to the Poisson law of rare events as

$$n = L(1 - e^{-Np/L})$$

where L equals the number of the phoneme group in question from Table 39, column 3, and p equals the probability of such phonemes from Table 40, column 5 (average). More precisely, L is the number of occasions on which a phoneme, belonging to a specified category, appears in phonological opposition, disregarding the possibility that in a number of cases the opposition may not be phonologically relevant (Trubetzkoy's "Aufhebungsstellung").

The argument implied in the above formula is as follows: according to the law of rare events, the probability of one of L phonological oppositions not occurring in a sample of one phoneme is $e^{-1/L}$; the probability of one of L phonological oppositions not occurring in a proportion Np of a text of length N is $e^{-Np/L}$, and consequently the probability of its occurrence is $1 - e^{-Np/L}$. The probable number of phonological oppositions of the group in a text length N is then $L(1 - e^{-Np/L})$.

The numerical data which form the basis of the calculations are listed in Table 41.

We see from Table 42 that the sample occurrence of 1000 consonant phonemes is accounted for to 96% and that of 5000 to 84% by the theoretically expected number

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 41

	Vocabulary (Table 39, col. 3) L	Occurrence Probability (Table 40, col. 5) p
Labials	3377	.235
Dentals	8296	.649
Palatals	207	.010
Velars	1982	.106
Total	13862	1.000

TABLE 42

Calculated frequencies for samples of different size

	$N = 1000$	$N = 5000$
Labials	228.3 (23.6%)	992.8 (23.6%)
Dentals	622.0 (64.3%)	2687.9 (64.1%)
Palatals	9.7 (1.0%)	44.5 (1.1%)
Velars	107.0 (11.1%)	565.8 (11.1%)
Totals	967.0 (100.0%)	4191 (100.0%)

of phonemes, using a law of chance. This means that the number of different consonantal phonological oppositions in the text samples are, to that extent, accounted for by those in the dictionary, representing the functional burdening of phonemes.

The remaining 4% and 16% are repetitions. Moreover, there is a striking agreement between the actually observed relative frequencies in running texts and those calculated from the dictionary probabilities, using a law of chance, in the proportion of phonemes accounted for by the four phoneme groups (cf. Table 40, col. 5 and Table 42).

8.4 CONCLUSIONS

Functional burdening thus appears as the dominant factor in the use of phonemes in speech, and the deviation of the frequencies of certain phoneme groups in speech from those in the dictionary (dentals and velars) is not an argument against the derivation of the former by a chance mechanism from the latter. The mutual relation between phonemes as regards functional burdening determines that of the categories of phonemes in speech output. This is what gives a language its specific character and makes us at once appraise the difference between two languages when heard spoken, even though we may not understand a word of either. The stability of distribution of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

functional burdening among the phonemes of a language, and the characteristic differences in this respect between nations is reminiscent of other racial characteristics, e.g. blood-group distributions. For purposes of comparison, Table 43 gives phonemic distributions for five languages (Krámský, 1956).

TABLE 43

		Engl. %	Ital. %	Pers. %	Turk. %	Hung. %
Manner of articulation	Plosives	41.4	42.36	29.8	36.19	39.23
	Fricatives	21.6	16.74	39.1	31.47	25.05
	Nasals	37.3	40.90	31.1	32.34	35.72
Place of articulation	Labials	24.4	20.67	20.26	18.05	17.50
	Dentals	59.8	62.84	48.37	60.20	57.64
	Palatals	1.5	6.25	10.59	4.32	9.19
	Velars	14.3	9.97	20.78	17.43	15.69

Our results admit of the conclusion that the distribution of phonemes according to mode and place of production when used in speech can be regarded as the phenotype of their functional burdening, and with it of the phonemic structure as the genotype of the language, in the same way as other parts of the phenotype of the human species can be regarded as manifestations of the genotype. In other words, the utilisation of a phoneme group in speech is closely similar to its functional burdening in the dictionary or in "la langue". This means that Trubetzkoy's "double relativity" is an illusion in this case: the functional burdening of a phoneme group in use being only a random sample of that in the dictionary.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

IX

THE RELATION BETWEEN THE VOCABULARY DISTRIBUTION AND THE OCCURRENCE DISTRIBUTION OF WORD LENGTH

One of the many unsolved problems of quantitative linguistics is the relation between the vocabulary and the occurrence distribution of specified linguistic forms, i.e. between the frequency distribution in the dictionary and that of occurrence in the spoken or written language.

For different linguistic forms, such as phonemes, phoneme combinations (morphemes), word length (in terms of syllable, phoneme, letter number), the answer may conceivably be different. The present investigation deals with the characteristic of word length in terms of phoneme and letter number per word, and arrives at the conclusion that the occurrence distribution can be regarded as a moment distribution of the vocabulary distribution. This will be shown to be a consequence of the lognormality of word length distributions.

To lognormality of certain linguistic distributions as an empirical fact Williams has drawn attention (1940, 1946). In this chapter the hypothesis of lognormality is extended to a set of related linguistic variables, viz. the distributions of both word occurrence and vocabulary according to word length in terms of the number of letters as well as phonemes. The comparison of these distributions as lognormal variates reveals the hypothesis of lognormality as being of great value for the study of quantitative linguistics, practical and theoretical.

9.1 LOGNORMALITY OF THE WORD LENGTH DISTRIBUTION

The material for the investigations was provided by a count of approximately 80,000 word occurrences obtained from telephone conversations by French, Carter and Koenig of the Bell Telephone Co. (1930). The distributions according to word length of the 76,054 word occurrences and of the 738 vocabulary items, that is different words contained in the sample, in terms of both letters and phonemes are shown in the following table (after Tables 55, 56 Vol. 1), where p_i is the percentage of words containing i units and x_i the average frequency of such words in the spoken language.

As shown below, if the occurrence distribution is a moment distribution of the vocabulary distribution, their standard deviations must be sensibly equal, which will

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 44

No. of units i per word	Letters				Phonemes			
	Vocabulary		Occurrence		Vocabulary		Occurrence	
	p_i	$\sum_{i=1}^{k} p_i$	$p_i x_i$	$\sum_{i=1}^{k} p_i x_i$	p_i	$\sum_{i=1}^{k} p_i$	$p_i x_i$	$\sum_{i=1}^{k} p_i x_i$
1	0.3	0.3	8.0	8.0	0.4	0.4	8.0	8.0
2	3.2	3.5	20.9	28.9	8.1	8.5	36.8	44.8
3	9.8	13.3	25.2	54.1	30.8	39.23	35.0	79.8
4	26.5	39.8	26.2	80.3	23.4	62.7	11.4	91.2
5	19.1	58.9	9.45	89.8	14.6	77.3	4.4	95.7
6	15.6	74.5	4.36	94.1	8.0	85.3	1.6	97.2
7	10.0	84.5	2.5	96.6	6.0	91.3	1.3	98.5
8	6.1	90.6	1.87	98.5	4.5	95.8	0.8	99.3
9	4.2	94.8	0.8	99.3	2.7	98.4	0.5	99.8
10	2.6	97.4	0.46	99.8	1.1	99.5	0.08	99.96
11	1.9	99.3	0.25	99.05	0.14	99.7	0.01	99.98
12	0.4	99.7	0.03	99.99	0.27	100.0	0.02	100.0
13	0.2	100.0	0.01	100.0	—	—	—	—
Mean of $\log_{10} i$	0.703		0.494		0.608		0.414	
S.D. of $\log_{10} i$	0.168		0.218		0.189		0.187	

result in the parallelism of the lines representing these distributions on a logarithmic probability grid. For the phoneme distribution this is the case, for the letter distribution they differ by 0.050, the lines being slightly convergent.

The plot of these distributions on a log probability grid whose abscissa has a logarithmic scale and whose ordinate is the Gaussian integral, is in each case a sensibly straight line, which suggests lognormality of the variates (Fig. 11). If the hypothesis of lognormality is correct, they should comply with the characteristic features of such distributions (Aitchison & Brown, 1957). As the most characteristic feature—and peculiar to lognormal variates[1]—we take the moment distribution property, according to which the j-th moment distribution of a lognormal distribution with parameters γ and σ^2 is also a lognormal distribution with parameters $\gamma + j\sigma^2$ and σ^2 respectively.

Writing for the logarithmic distribution function $\Lambda(x \mid \mu, \sigma^2)$, where μ is the logarithmic mean and σ the logarithmic standard deviation, we define the j-th moment distribution function as

$$\Lambda_j(x \mid \mu, \sigma^2) = \frac{1}{\lambda_j'} \int_0^x u^j \, d\Lambda(u \mid \mu, \sigma^2), \tag{3}$$

where $\lambda_j' = e^{j\mu + \frac{1}{2}j\sigma^2}$ is the j-th moment about zero. For the proof, see appendix A.

It follows that the graphs of different moment distributions should be parallel lines on log probability paper, with a distance of $j\sigma^2$ between them.

[1] Though not exclusively so.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

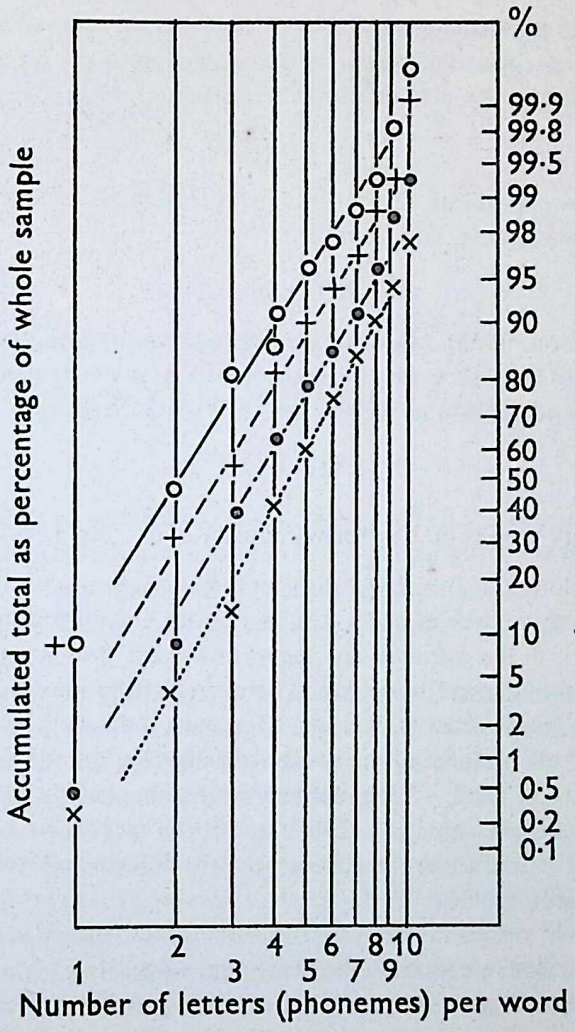


FIG. 11

9.2 DISTRIBUTION OF OCCURRENCES AS MOMENT DISTRIBUTION OF VOCABULARY DISTRIBUTION

Turning now to the distributions of Table 44 and Fig. 11, we find the logarithmic probability graphs for vocabulary and occurrences for both letters and phonemes, to be sensibly parallel lines. If what is plotted are different moment distributions, the distance between the vocabulary and occurrence lines must be that required by the moment theorem.

What is plotted against log word length for vocabulary is $\sum_{i=1}^{k} p_i$, and for occurrence $\sum_{i=1}^{k} p_i x_i$, where p_i is the number of vocabulary items (different words) of length i, and x_i

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the average number of occurrences of a word of length i. In an efficient code, such as language may be regarded to represent, word length appears to be inversely related to frequency of occurrence by a function of the form

$$f(i) = (a/i^k)\, b^i \tag{4}$$

where a, b and k are constants, and where b is so close to unity that as a first approximation the formula may be written as

$$f(i) = ai^{-k} \tag{5}$$

(Good, 1951; Simon, 1955; Mandelbrot, 1954). In our notation $x_i = ai^{-k}$. If i stands for length of English words in terms of letters or phonemes, k appears to be between 2 and 3, and we assume it to be 2.4. We write therefore

$$x_i \propto i^{-2.4} \tag{6}$$

The log probability graphs of $\sum_{i=1}^{k} p_i$ for vocabulary and $\sum_{i=1}^{k} p_i x_i \left(\propto \sum_{i=1}^{k} p_i\, i^{-2.4} \right)$ for occurrences plotted against log i thus represent respectively the basic frequency distribution and the -2.4th moment distribution for vocabulary items according to log word length. If the variates are lognormal these graphs may, therefore, be expected to show the characteristic relation between distributions of different moments explained above. This is actually the case. Moreover, if the distribution of occurrence represents the -2.4th moment of the vocabulary distribution, the averages of the two distributions, as the 0th and -2.4th moment means, should be $2.4\sigma^2$ apart, that is the occurrence mean should equal $\mu - 2.4\sigma^2$; as will be seen, there is good agreement between observation and theory in this respect, the differences between the respective means being sensibly equal to $2.4\sigma^2$.

Table 45 gives in column 3 the observed difference between the vocabulary and occurrence means, and in column 6 the theoretical value.

The theoretical difference between the moment averages as given in the text is in terms of natural logarithms. If logarithms to the base 10 are used, the following trans-

TABLE 45

	Logarithmic mean (base 10)	Difference	Logarithmic S.D. (base 10)	Average S.D. ($\bar{s}$)	$-2.4 \times 2.3026 \times \bar{s}^2$
Phonemes					
Vocabulary	0.608	—0.194	0.187	0.188	—0.195
Occurrences	0.414		0.189		
Letters					
Vocabulary	0.703	—0.209	0.168	0.193	—0.206
Occurrences	0.494		0.218		

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

formation must be observed (Herdan, 1953). Since the mean and standard deviation of the log distribution are the logarithms of the geometric mean G and the geometric standard deviation σ_G, respectively, viz.

$$\mu = \ln G, \ \sigma = \ln \sigma_G \tag{7}$$

the relation

$$\mu_j = \mu + j\sigma^2 \tag{8}$$

can be written as

$$\ln G_j = \ln G + j \ln^2 \sigma_G \tag{9}$$

Transforming into logarithms to the base 10, we have

$$\log_{10} G_j = \log_{10} G + 2.3026\, j \log_{10}{}^2\sigma_G$$

or

$$\mu_j = \mu + 2.3026\, j\sigma^2 \tag{10}$$

The last column of Table 45 shows $2.3026\, j\sigma^2$ with $j = -2.4$ and with the mean values, $\bar{s}$, from the preceding column substituted for σ. These figures are in good agreement with the observed differences given in the third column.

We may, therefore, take it as a reasonable hypothesis supported by some statistical evidence that:

(a) the distribution of vocabulary and that of occurrence against word length in terms of number of letters or phonemes satisfy the criteria for lognormality;

(b) the distribution of occurrences may be regarded as the -2.4th moment distribution of vocabulary according to word length.

The practical importance of this lies in the fact that knowing either the vocabulary or the occurrence distribution, it is possible to obtain immediately an estimate of the other distribution. Let us assume we had made a vocabulary count and wanted an estimate of the distribution of occurrences of that vocabulary in a continuous text of a certain length. Knowing the logarithmic mean and the logarithmic standard deviation σ of the vocabulary count, we calculate the average number of occurrences as $\mu - 2.4\sigma^2$, and knowing that the distributions are lognormal with the same standard deviation, we can immediately construct the straight line on log probability paper which represents the cumulative distribution of occurrences.[1] Working backwards from this, we obtain the distribution function. This may be of interest in connection with mechanical translation. Given a vocabulary with a certain distribution of words according to their length, we can quickly obtain an estimate of the distribution of occurrences in a text of a given size.

Considering that the method rests upon the assumption of an inverse relation between frequency and word length, which may be true to a different degree for different material, it follows that the multiple of 2.4 will give a satisfactory solution only to the extent to which the assumption is in agreement with the facts. The less true

[1] Parallel to the original line for the vocabulary distribution with a distance of $-2.4\sigma^2$ between the lines.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the assumption of strict reciprocity, the more will the multiple differ from 2.4 and it will, therefore, in general, be necessary to ascertain by a pilot investigation or from the literature how far the reciprocity assumption is true, in order to choose the correct multiple of σ^2. For determining the constant, Mandelbrot's (1953) or Simon's (1955) formula will be found useful.

Conversely, from the distance of the log probability graphs for vocabulary and occurrence distribution, inferences may be drawn about the extent to which the reciprocity relation (Zipf, 1949; Simon, 1955) is true for a given language.

9.3 THE RELATION BETWEEN OCCURRENCE AND LEXICON DISTRIBUTION IN THE LIGHT OF INFORMATION THEORY

It is of interest to compare the two distributions with regard to the information they contain. Considering that in an efficient coding system, $\log i \propto \log 1/p_i$ (p_i standing for the probability of occurrences of symbols of length i) and $\Sigma p_i \log i \propto \Sigma p_i \log 1/p_i$, the first moments of our distributions appear to be proportional to the information I_v for vocabulary and I_o for occurrences which, according to the theory of information (Brillouin, 1956), are calculated as

$$I_v = -K \Sigma p_{vi} \log p_{vi}$$

and (11)

$$I_o = -K \Sigma p_{oi} \log p_{oi}$$

where p_{vi} is the probability of word length i in the dictionary and p_{oi} in the spoken language. For a binary system, for which $K = 3.322$, these quantities become in our case:

Letters: $I_v = 2.988$ bits/word length, $M = 5.417$
$I_o = 2.628$ bits/word length, $M = 3.508$

Phonemes: $I_o = 2.726$ bits/word length, $M = 4.413$
$I_o = 2.274$ bits/word length, $M = 2.856$ (12)

This enables us to compare the information property of letters against phonemes, and for each of vocabulary against occurrences. What we are doing, in using the entropy I as a characteristic instead of the average M, is to make use of the fundamental rule which plays a most important part in statistical thermodynamics, viz. that the average and the most probable distribution coincide for large numbers. The relation between the two is for binary coding expressed by Shannon's fundamental theorem $I \leqslant M$.

When trying to assess which code is more efficient, we must not be misled by the impression that a greater I means more information, and thus a more efficient code. The entropy I which originally was considered by Shannon as a measure of information is now recognized (by Brillouin, for instance) as being a measure of unexpectedness or lack of information, which has compelled Brillouin to introduce the concept of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

negentropy for measuring information. The apparent confusion of thought is resolved by the following consideration: a small entropy (and a great redundancy) implies a small effort in terms of number of guesses for restoring missing information; only a great store of information about the structure of the code will enable the receiver of the message to make correct guesses with little effort, and a small I (great R) thus implies much *past* or *advance* information. On the other hand, if we think in terms of what is *gained* by such guesses, that is, *future* information, then it is clear that symbols with great probability of occurrence which are easily guessed correctly will add little to our knowledge, and thus represent little information, whereas rarely occurring symbols which are only guessed with difficulty mean much information (and little redundancy).

For our data we find

$$\text{(a)} \qquad \begin{aligned} I_v(\text{phon.}) &= 2.726 < I_v(\text{lett.}) = 2.988 \\ I_o(\text{phon.}) &= 2.274 < I_o(\text{lett.}) = 2.628 \end{aligned} \tag{13}$$

and thus $I(\text{phon.}) < I(\text{lett.})$

Further we find

$$\text{(b)} \qquad \begin{aligned} I_o(\text{phon.}) &= 2.274 < I_v(\text{phon.}) = 2.726 \\ I_o(\text{lett.}) &= 2.628 < I_v(\text{lett.}) = 2.988 \end{aligned} \tag{14}$$

and thus $I_o < I_v$.

This means

(a) that the phonemic channel is more efficient than the letter channel; and

(b) that the knowledge of occurrences makes it easier to guess at missing information and that it therefore represents some additional information over and above that provided by the mere vocabulary code.

The basis of linguistic information theory is the hypothetical solidarity between the various items of a specified linguistic characteristic, word length in our case. Such solidarity exists between the items in the dictionary and between their frequency of occurrence in the spoken and written language, and makes it possible to guess—with some expectation of being correct—at the word length of individual items chosen at random from either dictionary or speech. The number of such guesses for arriving at a probably correct result is epitomized by the value of the entropy. The fact that I_o is less than I_v means that the uncertainty of word length in speech is less than in the dictionary. This is only what one might expect, since in speech the solidarity of word length in the dictionary is supplemented by that of occurrence, and we have therefore more information available for guessing word length correctly.

Returning now to the conclusion reached in Sect. 9.1, 9.2, it seems rather interesting that the introduction of the lognormal hypothesis for the mathematical structure of our distributions should enable us to add, and subtract, information when only one distribution, vocabulary or occurrences, is given. For this is what the estimation of the other distribution means in terms of information theory. Being given the vocabulary distribution and deriving that of occurrences by the method established above,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

we reduce I_v to I_o which means less deciphering work because of more advance information, and vice versa if the vocabulary distribution is derived from the occurrence distribution.

Summary

1. From a representative sample of 76,054 words of spoken English the frequency distributions of vocabulary and occurrences according to word length in terms of letters and phonemes were obtained, and subjected to statistical and information-theoretical analysis.

2. The distributions were found to be sensibly lognormal.

3. The relation between the occurrence and vocabulary distributions according to either letter or phoneme number is such that they could be interpreted as different moment distributions of a lognormal variate. This admits the derivation of one distribution by the moment distribution theorem for lognormal variates if the other is known. The practical importance of this lies in the possibility of a quick estimate of the occurrence distribution without having to carry out a complete word count, or of the vocabulary distribution without a vocabulary count.

4. A transformation of the logarithmic variable of the distributions by using an established relation between frequency and length of linguistic symbols shows the occurrence and vocabulary means to be of the form of the theoretical measure of information. The comparison of the statistical results with their information-theoretical interpretation enables us to assess the value, in terms of information, of using the moment distribution theorem of lognormality.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

X

SELECTIVE VARIATION IN THE VOWEL-CONSONANT SEQUENCE OF ENGLISH AND RUSSIAN

The method described in this chapter has a double purpose:

(a) to find a numerical expression for selective variation in the vowel-consonant sequence in a given language, and to account for it in terms of the particular transition probabilities between vowels and consonants, and their combinations in pairs, triplets, etc., and

(b) to compare selective variation in vowel-consonant sequences in different languages, and in particular in English and Russian.

The method is that of the Markoff chain, so called after the Russian mathematician, A. A. Markoff, whose exposition of his theory in the *Bulletin of the Imperial Academy of Sciences of St. Petersburg* in 1913 may be regarded as one of the classical papers of mathematical linguistics, though few linguists will have a first-hand knowledge of it. Markoff's purpose, however, was not to contribute to the science of linguistics, but to give an application of his statistical theory of chains. Consequently, although faultless as a piece of applied mathematics, his paper is open to certain criticism from the linguistic point of view. Apart from dealing with letters and not with phonemes, the method requires a continuous sequence of linguistic elements disregarding the empty spaces between words, with the consequence that what is, for instance, called "vowel sequence" does not only refer to the vowel sequence within words, but comprises also that between the end and beginning of words. However, I have followed his method closely, for these reasons. Statistical linguistics in their present state suffer from one disadvantage which was already stressed by Trubetzkoy in 1939, and that is their lack of uniformity which prevents valid comparisons. A typology of languages requiring statistical methods is impossible as long as results are not comparable. It is readily admitted that very often the discrepancies arise from the desire on the part of the investigator to correct or improve upon methods previously used, but it should be made an important methodological point to weigh the advantage of a change in experimental design against that of uniformity. In the present case I thought it best to design the experiment in accordance with that carried out by Markoff, and thus ensure the possibility of valid comparisons between results, being fully aware of the limits imposed by Markoff's particular design, viz. that the comparison is restricted to one of *printed English and Russian* and that the vowel-consonant sequences referred

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

to are not only those within words, but those within and between words. Within those limits, however, the procedure is not only unobjectionable, but may serve as a model for other typological investigations.

10.1 MATERIAL AND METHOD

Markoff's material were the 20,000 letters comprised in the first chapter and the first 16 Sonnets of the second chapter of Pushkin's novel in verse *Evgeny Onegin*. In order to make the comparison between English and Russian letter sequences a truly valid one, and to forestall criticism on account of difference in subject matter and style, the first 10,000 letters, representing the first 29 sonnets and the first 4 lines of the 30th, of the English translation of Pushkin's poem by Oliver Elton were chosen as the material of the present investigation. This gives a sequence of 10,000 connected observations, each of which is either a vowel or a consonant. From this material, it is possible to obtain an estimate of the basic probability of a vowel, or a consonant, in continuous pieces of the English language; we shall denote these probabilities by p and $q = 1 - p$ respectively. Moreover, we can obtain the estimates of the probability of a vowel following a vowel $-p_1$, of a vowel following a consonant $-p_0$, of a vowel following two vowels $-p_{1,1}$, of a vowel following a consonant which is preceded by a vowel $-p_{1,0}$, of a vowel following a vowel which is preceded by a consonant $-p_{0,1}$, and a vowel following two consonants $-p_{0,0}$. The complementary probabilities are denoted respectively by q_1, q_0 etc.

10.2 CONNECTED AND DISCONNECTED LETTER OCCURRENCE

The total sequence of 10,000 letters is divided into 100 groups of 100 letters each, as they followed each other in the text. With a view to combining the observations in two ways, which is necessary for our purpose, we arrange every group of 100 successive letters in a square as follows:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
.....									
91	92	93	94	95	96	97	98	99	100

and count in each column the number of vowels. We then combine the counts in columns

1st & 6th, 2nd & 7th, 3rd & 8th, 4th & 9th, 5th & 10th

and denote these sums by the symbols

(1,6), (2,7), (3,8), (4,9), (5,10).

This gives five figures for each text sample of 100 whose sum

$$(1,6) + (2,7) + (3,8) + (4,9) + (5,10)$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

is the total number of vowels in a sequence of 100 letters. In this way we obtain 100 figures for the vowel number in successive samples of 100 letters.

If now, instead of summing over the different column-pairs belonging to one sample of 100 consecutive letters, the double-column counts each representing the number of vowels per 20 letters are combined for 5 samples of 100 letters each by summing over the *same* column-pairs in the 5 samples, we get a new distribution of the 10,000 letters in groups of 100, and the quantities

$$\Sigma(1,6),\ \Sigma(2,7),\ \Sigma(3,8),\ \Sigma(4,9),\ \Sigma(5,10)$$

represent new figures for the number of vowels in samples of 100 letters, viz. the vowel number in disconnected but corresponding column-pairs.

Thus, whereas the first summation gives the vowel numbers in successive or connected parts of the text, the second gives the vowel numbers in disconnected parts of the

TABLE 46

7	8	8	12	11	46	13	9	7	7	10	46	12	12	9	9	10	52	8	12	10	7	7	44
8	4	12	2	9	35	7	8	9	7	8	39	6	5	9	8	7	35	7	6	7	7	8	35
4	11	3	7	7	32	8	7	11	7	8	41	7	7	9	8	9	40	13	7	6	10	7	43
10	8	9	9	5	41	8	12	4	10	12	46	4	10	7	7	8	36	7	8	10	10	10	45
8	9	6	7	8	38	5	6	10	8	6	35	6	7	5	4	9	38	5	3	5	9	10	32
37	40	38	37	40	192	41	42	41	39	44	207	35	41	39	36	43	194	40	36	38	43	42	199
8	12	8	6	5	39	9	9	6	6	12	42	8	9	7	3	11	43	6	10	9	7	7	39
8	5	5	8	10	37	6	8	8	6	9	37	6	7	7	11	7	38	7	6	6	11	9	39
7	9	11	8	7	42	9	10	6	8	5	38	7	8	9	3	5	32	7	10	8	4	6	35
8	7	7	10	3	40	2	9	11	8	12	43	12	10	5	8	11	46	7	9	9	9	9	43
10	6	10	11	7	44	9	8	7	10	7	41	9	4	11	6	7	37	10	5	6	6	12	39
41	39	41	43	38	207	35	44	38	38	46	201	42	38	39	36	41	196	37	40	38	37	43	195
6	8	7	6	5	32	6	8	7	9	9	39	9	7	12	12	6	46	10	12	11	8	6	47
7	6	9	10	11	43	6	8	7	3	3	27	7	10	3	8	4	32	5	6	10	10	10	41
7	10	5	9	6	37	12	8	7	8	10	45	8	8	11	6	11	44	8	7	6	7	9	37
9	8	9	8	8	42	6	7	12	7	6	38	7	10	13	9	10	49	10	12	9	9	6	46
7	9	8	7	8	39	8	10	5	9	7	39	4	8	5	4	7	28	9	6	9	9	7	40
36	41	38	40	38	193	38	41	38	36	35	188	35	43	44	39	38	199	42	43	45	43	38	211
10	6	8	7	6	37	9	6	9	6	7	37	6	6	9	7	10	38	8	7	7	5	5	32
9	6	7	4	9	35	10	8	8	6	7	39	9	5	6	7	3	30	7	6	10	8	5	36
6	8	5	6	7	32	3	11	6	7	5	32	8	11	6	10	11	46	8	6	4	7	11	36
7	8	9	9	7	40	10	5	4	8	9	36	3	7	12	6	7	35	8	6	9	9	4	36
4	8	9	8	9	38	3	7	12	10	9	41	10	8	9	10	7	44	11	9	4	10	10	44
36	36	38	34	38	182	35	37	39	37	37	185	36	37	42	40	38	193	42	34	34	39	35	184
9	10	7	8	7	41	5	10	11	2	10	38	7	9	7	6	7	36	10	12	10	12	8	52
6	5	8	12	7	38	11	5	7	8	7	38	5	6	6	8	10	35	11	2	9	13	8	43
12	14	9	5	8	48	9	11	9	7	7	43	9	8	7	5	8	37	4	10	13	6	9	42
7	7	5	8	10	37	5	9	9	8	9	40	4	6	6	7	7	30	10	12	7	5	9	43
9	6	8	7	5	35	9	8	7	10	5	39	8	9	9	7	6	39	7	5	7	7	6	32
43	42	37	40	37	199	39	43	43	35	38	198	33	38	35	33	38	177	42	41	46	43	40	212

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 47

Vowel number per 100 letters	Frequency of connected counts	Frequency of disconnected counts
27		1
28		1
29		
30		2
31		1
32		9
33	2	
34	3	
35	8	9
36	8	6
37	10	9
38	19	9
39	8	11
40	9	5
41	8	6
42	8	4
43	11	8
44	3	5
45	1	2
46	2	7
47		1
48		1
49		1
50		
51		
52		2
	100	100
Average	39.1	39.1
(Standard deviation)^2	6.73	29.11

sequence of 10,000 letters. From a difference between the two results, certain inferences may be drawn about the amount of selective variation present.

Table 46 exhibits the results of our vowel count in 20 blocks, each of which gives in the first line five figures for the (1,6) count and their sum, in the second five figures for the (2,7) count, etc., and, consequently, in the last line the vowel numbers for the first 100 letters of text, for the second 100 letters, etc., and, ultimately, the total vowel number in five groups of 100 consecutive letters. The last column gives the sums $\Sigma(1,6)$, $\Sigma(2,7)$, etc.

We concentrate first on the numbers in the last line of each square of Table 46, that is the vowel frequencies per 100 consecutive letters of text, from which we get the frequency distribution as shown in col. 2, Table 47.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The arithmetic mean results as

$$a = 39.07 \sim 39.1$$

The total number of vowels in the sample of 10,000 letters is 3907, and the total number of consonants 6093. Since a represents the average vowel number per 100 letters, the probability of a letter being a vowel is

$$39.1/100 = .391$$

which we shall denote by p. The complementary probability $1 - p = .609$ must be that of a letter being a consonant, and we shall denote it by q.

The square of the standard deviation, as a measure of dispersion, is calculated as the average square deviation from the mean, and results as

$$s^2 = 6.73$$

However, the standard deviation may be calculated in yet another way, namely under the assumption of a pure random or chance distribution of the vowel frequencies in groups of 100 letters. We then obtain for it, according to the so-called Bernoullian law

$$\sigma^2 = npq = 23.81 \qquad (15)$$

What we are interested in is whether there is a linkage, in the sense of selective affinity, between vowels and consonants, between vowels and vowels or consonants and consonants. The answer is obtained by comparing the observed standard deviation with the value that would arise on pure chance, that is without linkage. The ratio of the two standard deviations is known as Lexis' Q. If the fluctuations of the vowel frequencies in samples of 100 letters were completely random, there would be no significant difference between the observed and the theoretical measure of dispersion, and consequently the expectation of Q under this condition is unity.[1]

On the other hand, if the fluctuations of the vowel numbers cannot be regarded as random variations in a Bernoullian or normal statistical population of such numbers, then we should expect the ratio of the two standard deviations to differ significantly from unity. In the first case, we speak of a "normal" dispersion, and in the second of either a "super-normal" or a "sub-normal" dispersion, according to whether Q is greater or smaller than unity. A mass event with a constant probability leads to a Q indicative of normal dispersion, and a mass event with a variable probability to a Q indicative of super-normal dispersion, and since another possibility of unconnected chance events is unthinkable, it follows that a Q which is significantly less than unity does no longer refer to pure chance events, but must be indicative of a linkage or correlation between the individual events.

As every statistic obtained from a sample, Q has an error. This error is calculated as $2/(k-1)$, where k is the number of degrees of freedom or categories. We regard a

[1] Vol. I, Sects. 5.7, 18.6.

डा० विद्या निवास मिश्र
प्राध्यापक, भाषा विज्ञान
निजी पुस्तकालय
पुस्तक संख्या.....

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

value of Q as significantly different from unity, if it lies outside the range of twice or three times that error.

The ratio: observed/theoretical squared standard deviation, in symbols

$$Q = s^2/\sigma^2 \tag{16}$$

results for our data as

$$Q = 6.73/23.81 = .282$$

with a standard error of

$$2/14 - 1 = .153$$

Since Q is less than $1 - (3 \times .153) = .541$, it must be regarded as highly significant of a tie or linkage between the observations, that is between vowels and consonants.

10.3 THE MARKOFF CHAIN MECHANISM

In order to ascertain the type of linkage between vowels and consonants, we use the Markoff chain method. That is, we shall calculate a coefficient comparable to Lexis' Q, but on the basis of the actually observed numbers of joint occurrence of vowels and consonants. Agreement between the two coefficients, obtained independently, admits the conclusion that it is the peculiar mechanism of the Markoff chain which provides the explanation of the linkage ascertained in a global manner by the sub-normality of Lexis' Q.

Scanning the sequence of 10,000 letters for

vowel vowel

sequences, we find 696, which divided by the total number of vowels, 3907, gives

$$p_1 = 696/3907 = .178$$

In a similar manner, we could calculate the probability of

consonant consonant

sequences. However, it is possible to save ourselves the tedious counting labour by the following considerations. Subtracting 696 from 3907 we get 3211 as the number of consonants following a vowel, which is also the number of vowels following a consonant, and thus briefly, the number of joint vowel-consonant occurrences. Since now every consonant, except the first, follows either a vowel or a consonant, the difference between the total consonant number and the number of joint vowel-consonant occurrences

$$6093 - 3211 = 2882$$

must be the number of

consonant consonant

occurrences.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

We then get for the probability of a joint vowel-consonant occurrence

$$p_0 = 3211/6093 = .526$$

It should be observed that we use as the denominator for p_0, not the total vowel number, but the total consonant number. Since there are more consonants than vowels in our sample, we must expect, if the distribution were random, the joint vowel-consonant occurrences to exceed in number the vowel-vowel occurrences, and the comparison of the two numbers would tell us nothing about the linkage in which we are interested.

The difference $p_1 - p_0$, which is an index of the preference of joint vowel-consonant occurrence, we obtain as

$$\delta = .178 - .526 = -.348 \tag{17}$$

On the assumption that our sequence of 10,000 letters represents a "*simple Markoff chain*", the theoretical value of the coefficient of dispersion with which to compare the empirical value of Q is calculated as the quantity

$$\frac{1+\delta}{1-\delta} = \frac{.652}{1.348} = .483 \tag{18}$$

The difference of this value from unity is indicative, just like a sub-normal Q, of a bond or linkage between the events.

But although pointing in the same direction, the two values do not approximate to one another too closely. The reason may be that, so far, we have only considered a simple chain, or a chain of the first order, that is one of letter pairs. From an examination of the letter sequences, it is evident that not only double, but triple etc. combinations occur which must be considered if the sub-normality was to be accounted for more fully. Consequently, if we wish to achieve by the chain method a closer approximation to Q, we must consider the letter sequences in the light of what Markoff has called a "*complex chain*", i.e. a chain of the second or higher order.

We first scan the sequence of 10,000 letters for the triple combinations.

vowel vowel vowel

and

consonant consonant consonant

and obtain the first as 85, and the second as 969. Dividing these numbers by the vowel-vowel and consonant-consonant numbers, respectively, we get the probabilities

$$p_{1,1} = 85/696 = .122$$

and

$$q_{0,0} = 969/2882 = .336$$

Considering that a vowel-vowel combination (vv) may be preceded by either a vowel (v) or a consonant (c), and that, therefore, the vv occurrences may form part of either a

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

vvv or a *cvv* (*vvc*) occurrence, the difference $vv/v - vvv/vv$ must represent the probability of a *v* occurrence forming part of a *cvv* (*vvc*) occurrence. Similarly, the difference $cc/c - ccc/cc$ represents the probability of a *c* occurrence forming part of a *vcc*(*ccv*) occurrence. Dividing each result by the corresponding probability of simple joint occurrences *cv* (*vc*), we get

$$\frac{cvv/v}{cv/v} = cvv/cv \tag{19}$$

being the probability of a simple joint occurrence *cv* belonging to the triple combination *cvv*, and

$$\frac{vcc/c}{vc/c} = vcc/vc \tag{20}$$

being the probability of a simple joint occurrence *vc* belonging to the triple combination *vcc*.

These quantities are quite analogous to cv/v and vc/c, respectively.

Combining the probabilities obtained before,

$$\begin{aligned} p &= .391 & q &= .609 \\ p_1 &= .178 & q_1 &= .822 \\ p_0 &= .526 & q_0 &= .474 \\ p_{1,1} &= .122 & q_{0,0} &= .336 \end{aligned} \tag{21}$$

as indicated above, we calculate the triple transition probabilities between vowels and consonants as follows:

$$\varepsilon = \frac{vvv/vv - vv/v}{cv/v} = \frac{p_{1,1} - p_1}{q_1} = \frac{.122 - .178}{.822} = -\frac{.056}{.822} = -.068 \tag{22}$$

$$\eta = \frac{ccc/cc - cc/c}{cv/c} = \frac{q_{0,0} - q_0}{p_0} = \frac{.336 - .474}{.526} = -\frac{.138}{.526} = -.262 \tag{23}$$

We then calculate the coefficient of dispersion under the assumption of a complex Markoff chain as

$$\frac{1+\delta}{1-\delta}\left\{\frac{1+\varepsilon}{2(1-\varepsilon)} + \frac{1+\eta}{2(1-\eta)}\right\} + \frac{(q-p)(\eta-\varepsilon)}{(1-\varepsilon)(1-\eta)} \tag{24}$$

Substituting in the above formula the values for $\delta, \varepsilon, \eta, p, q$, we get

$$.483\,\frac{.466}{1.068} + \frac{.369}{1.262} + \frac{(.609 - .391)(-.262 + .068)}{.932 \times .738} = .319 \tag{25}$$

which approximates the value for Q (.288) satisfactorily. We could, of course, get a still closer approximation by considering fourfold, fivefold, etc. letter combinations.

The coefficient of dispersion which we calculated from the quantities

$$p, q, \delta, \varepsilon, \eta$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

under the assumption of a complex Markoff chain thus agrees satisfactorily with that obtained empirically by the method of Lexis' Q. We are therefore justified in regarding the particular selective affinities between vowels and consonants, which we call a complex Markoff chain, as the explanation of the sub-normal Lexis' Q. The results are collected in Table 48, together with the corresponding ones for the Russian letter count. In order to write the transition probabilities as positive quantities, the table lists the values $-\delta$, $-\varepsilon$, $-\eta$.

TABLE 48

Letters (v = vowel, c = consonant)	English (10,000 letters)	Russian (20,000 letters)
	Numbers	
v	3907	8638
c	6093	11361
vv	696	1104
cc	2882	3827
cv (vc)	3211	7534
vvv	85	115
ccc	969	505
	Probabilities	
$v/(v+c)$ .... p	.391	.432
$c/(v+c)$ .... q	.609	.568
vv/v .... p_1	.178	.128
cv/v .... q_1	.822	.872
cv/c (vc/c) .... p_0	.526	.663
cc/c .... q_0	.474	.337
vvv/vv .... $p_{1\cdot1}$	.122	.104
ccc/cc .... $q_{0\cdot0}$	.336	.132
$cv/c - vv/v$ .... $p_0 - p_1 = -\delta$	.348	.535
cvv/cv .... $\frac{p_1 - p_{1\cdot1}}{q_1} = -\varepsilon$	.068	.027
vcc/vc .... $\frac{q_0 - q_{0\cdot0}}{p_0} = -\eta$	.262	.309
Markoff simple chain coefficient	.483	.300
Markoff complex chain coefficient	.319	.195
Lexis' Q	.282	.208

We now turn to the second summation according to column-pairs of the same denomination, and thus to the vowel count in disconnected samples of 100 letters each. The frequency distribution of these numbers, which we denoted by $\Sigma(1,6)$, $\Sigma(2,7)$, etc. and listed as the last column of each block in Table 46, is given in Table 47, col. 3.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The arithmetic mean is, of course, the same as for the distribution of consecutive letters, viz.

$$a = 39.1$$

The standard deviation (squared) results as

$$s^2 = 29.11$$

From this we obtain Lexis' Q as

$$Q = 29.11/23.81 = 1.222$$

Since the error of Q is $2/(20—1) = .10$, the difference of Q from unity is still within three times its error, and thus need not be regarded as significant, which implies that the vowel-consonant sequence in disconnected parts of text is not essentially different from a random sequence. The difference between the squared standard deviations obtained from connected pieces of text, 6.73, and from disconnected parts, 29.11, is sufficiently great to convince us of the presence of bonds or affinities between the individual events in the former.

10.4 COMPARISON OF RUSSIAN AND ENGLISH VOWEL-CONSONANT SEQUENCES

(a) Comparing the vowel-consonant sequences in printed English and Russian, we find it non-random in both languages, but subject to selective variation, due to selective affinities between vowels and consonants. The resulting structure corresponds to what mathematicians call the Markoff chain.

The implied bond or correlation between the types of letters, which is significant in both languages, is measured by the deviation of Lexis' Q from unity. That this difference is slightly greater for Russian, does not necessarily mean that the bond between the types of letters is stronger there, since a sub-normal Q decreases with sample size until a stable value is reached. Other things being equal, we should, therefore, expect it to be smaller for the Russian count of 20,000 letters, than for the English one of 10,000. In any case, it is not so much the difference between the values for Q which interests us, as the differences between the particular probabilities and transition probabilities by which to account for the deviation of Q from unity.

(b) Turning now to these probabilities, and the differences between corresponding ones in English and Russian, it should be noted that, unlike Q, they are not affected by the difference in size of the English and Russian samples. The count of English letters was stopped at 10,000 because it was found that this was a sufficient sample size for the different probabilities to approach their true values closely enough to be regarded as valid estimates of the true values (in the sense of the Law of Great Numbers).

The probability, q, of English consonants is greater than that of Russian consonants: consequently, that of English vowels, p, is smaller. On the other hand, both vowel and consonant pairs (vv and cc) have a greater probability in English. The reverse is true for the vc (cv) combination which has a greater probability in Russian. These results

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

would seem to point to English preferring letter clusters of the same type (*vv* and *cc*), and Russian those of the mixed type (*vc*~*cv*). This shows that the difference between Q_R and Q_E is, to some extent, systematic and indicative of the stronger tie between *v* and *c* in Russian.

This is confirmed by the higher probabilities of the *vvv* and *ccc* combinations in English, and the higher probability of the *vcc* combination in Russian; the *cvv* combination, on the other hand, has a greater probability in English.

(c) The difference between Q and the coefficient of dispersion obtained from δ is an indication that the chain is not a simple or first order one, i.e. one in which the bond does not extend beyond the immediate neighbour.

This is confirmed by the fact that the complex chain coefficient, which takes triple combinations into account, is a much better approximation to Q. This is true for both English and Russian, and shows the assumption of a complex Markoff chain structure to be correct for both languages. To be a little more explicit, it means that the selectivity of letters does not lead to mere repetition of either *v* or *c*, nor only to simple alternation of vowels and consonants, but that the selection extends to letters other than the immediate neighbour, resulting in a complex pattern of the vowel-consonant sequence.

(d) The value of Lexis' Q for the distribution of vowel frequency per 100 letters in disconnected parts of the sequence of 10,000 letters, namely in 5 column-pairs of the same denomination, is not significantly different, by the three-standard-error criterion, from the theoretical value according to the Bernoullian law. The dispersion is, therefore, not significantly different from normal, which implies a random sequence of vowels and consonants. This is what might have been expected, since by the method of summing over disconnected parts of the text, the ties between types of letters were severed. It also confirms the conclusion that the sub-normality of the distribution obtained by the first summation is due to the presence of ties between the types of letters.

(e) Apart from providing information for a typology of languages (Russian and English, in our case), as regards the vowel-consonant sequence, our investigation reveals in a rather striking way the inadequacy of the English alphabet as a written code for the English phonemic system.

Attention is first drawn to the fact that English uses more consonant letters than Russian, which as a Slav language is much richer in consonant phonemes. The same applies to consonant pairs which have a greater probability in English, and still more to consonant triplets, whose probability in English is almost three times that in Russian. That a language like English, which is rich in vowel phonemes and very economical in the use of consonant phonemes should, in the written language, be excessively burdened by consonant letters and their double- and triple combinations, and more so than Russian, which is rather poor in vowel phonemes, but very rich in consonant phonemes, shows how badly suited our alphabet is to the phonemic structure of the language.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Summary

1. The material for the investigation is a sequence of 10,000 letters representing the first 30 sonnets of the English translation by O. Elton of Pushkin's novel in verse *Evgeny Onegin*. The object of the investigation was to express numerically the selective variation in the sequence of vowels and consonants in a continuous text, and to compare the results with those obtained by A. A. Markoff from a count of the first 20,000 letters of the Russian original.

2. The vowel-consonant sequence in English is shown to be in accordance with the stochastic mechanism known as a complex Markoff chain. The numerical expression of selectivity so obtained is not only a yes-no answer (selectivity or randomness), nor is it restricted to giving only a global measure (in terms of the ratio Q, or rather its deviation from unity), but it analyses that empirical, global measure into its constituent elements in terms of the transition probabilities between vowels and consonants, each occurring singly, in pairs, triplets, etc.

3. On the basis of these results, English and Russian are compared with a view to arriving at a typology for the vowel-consonant sequence in these languages.

4. The comparison shows English, which as a spoken language is rich in vowel phonemes and very economical in consonant phonemes, as using more consonants and consonant clusters in alphabetic writing than Russian, which as a Slav language is notoriously rich in consonant phonemes and their combinations. This paradoxical result may be of interest to advocates of a spelling and alphabet reform for English, such as envisaged in G. B. Shaw's Will.

10.5 MARKOFF CHAIN AND MECHANICAL TRANSLATION

The type of linkage between linguistic elements described in this chapter as the Markoff chain, is of great importance in machine translation. Machine translation engineers have so far not been aware of this and, consequently, have not described their problems in terms of the Markoff chain, which may partly account for the lack of systematic presentation in that field.

How to use the ideas of this chapter in the context of machine translation problems will be shown in connection with a recent investigation by A. Koutsoudas (*Language and Speech*, I, Part IV, 1958, 265). The problem was how to make the machine identify the grammatical number of a Russian noun whenever form alone failed to provide the distinction. There are three groups of nouns having identical suffixes in the singular and plural. These are:

I. Feminine, inanimate, palatal and non-palatal nouns ending in -a and я.
Examples: газета = newspaper; миля = mile.

II. Feminine, inanimate, palatal nouns ending in -ь and -ия.
Examples: кость = bone; профессия = profession.

III. Neuter, inanimate, palatal and non-palatal nouns ending in -o, -ие and -е.
Examples: вино = wine; выражение = expression; море = sea.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Specimens of these groups found in a text of 15,000 running words were listed alphabetically together with the three preceding and three following words. The author then gives 10 rules for the machine to choose the plural translation. A form belonging to one of the groups I, II, III, is to be taken by the machine as being in the plural if it is preceded by one of the linguistic forms under rules 1-10. This implies a tendency of the two types of form to occur together, or in terms of the Markoff chain theory, a linkage of the Markoff type. In terms of general probability theory we would say that the forms under I-III and the preceding forms according to rules 1-10 occur jointly more often than could be expected on pure chance if the nouns in question are in the plural.

The advantage of expressing the method in this way is two-fold:

1. It suggests the need for ascertaining numerically the probabilities of joint occurrence for each of the indicator forms 1-10 with any of the nouns in groups I-III if in the plural, and further their comparison with the probabilities of joint occurrence by mere chance. Only if the former exceed the latter significantly, can the forms under rules 1-10 be regarded as reliable pointers to the plural translation of the nouns in groups I-III.

2. It explains and makes us realise—a matter only too often missed by machine translation engineers—that a bond such as that between the preceding forms (according to rules 1-10) and the plural translation of nouns in the groups I-III, is by no means a 100 per cent valid one, but only a probability link, and that a machine can, therefore, not be expected to give always the correct translation, but only with a certain probability. This makes us realise the unavoidable imperfection of machine translation. If even grammar rules are not 100 per cent perfect, but have a greater or smaller number of exceptions, how much less perfect must be the rules based only upon the linear sequence of language forms?

On the whole, it may be said that the realisation that machine translation problems of the type discussed here are mathematically of the nature of Markoff chains, will contribute to treating the whole matter more systematically, and avoid the sometimes haphazard, if not anecdotal, features of much of the work in that line.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XI

THE REPETITIVENESS OF LINGUISTIC FORMS

Languages are systems of coding, and since there must be fewer coding symbols (signs) than objects in the manifold to be denoted by them, the symbols will be used repeatedly, though with different frequencies, according to their usefulness.

The problem of sameness vs. otherness of linguistic forms or of diversity of expression, can be tackled in two ways. We may, while exploring the matter, preserve the linear order of the units, say words; or we may consider the text to be just a mass of words. In the first case, we study the spacing of particular words, or other linguistic forms, in the second we are interested in the repetitiveness of such units in the text regardless of their sequence. The statistical quantity used for describing the spacing of particular linguistic forms is *the distribution of gaps between repetitions*, or shortly, the gap distribution; that expressing repetitiveness in a mass of linguistic forms is the *repeat rate*.

11.1 THE DISTRIBUTION OF GAPS BETWEEN THE REPETITIONS OF LINGUISTIC FORMS

In what follows we shall consider the gaps between repetitions of words in terms of word number, but the method is fully capable of generalisation both as regards the units repeated (they may not only be words, but morphemes, phonemes, phoneme combinations etc., and on the other side, word categories, sentence types etc.), and as regards the units in terms of which gap length is expressed (they may be phonemes, letters, sentences, pages etc.).

We consider a word occurring several times in a text, or, as we say, the distribution of repetitions of a word. In the analysis of intervals between repetitions of a word, the text must be regarded as a sequence of places—sites—to be occupied by words. Hence the length of an interval between two occurrences of a certain word is measured by the number of such sites in the linear sequence between the occurrences in question, exclusive of both occurrences. Thus, if two successive occurrences are the same word, we say the interval length is zero; if they are separated by one site we say the interval length is 1, etc.

We are interested in the numerical distribution of intervals with respect to length in terms of the number of sites, that is in our case the distribution of intervals between

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

repetitions of words with regard to interval length in terms of the number of intervening words. However, as mentioned above, the unit of interval length may be lines or pages etc. The numerical distribution of intervals can be represented geometrically in a co-ordinate system by plotting the numbers of intervals of a given length as the ordinate against interval length as the abscissa. Instead of the numbers, we may plot —and it is recommended to do so for purposes of comparison—the relative frequency, that is, the number divided by the total number of sites. We then get the distribution curve of gap length.

The empirical distribution curve of gap length obtained in this way may be examined with regard to whether it is in conformity with any known distribution law. In particular, if we are interested in the question of determination vs. chance in the use of certain words, we would compare the observed distribution with a chance distribution of gaps. One such chance distribution is derived as follows.[1] We denote by

n ..... the total number of sites in the text in question, i.e. the total number of words,
f ..... the number of occurrences of the element in question, i.e. the number of occurrences of a particular word,
$p = f/n$ the relative frequency, and in great samples, an estimate of the probability of a particular word,
v ..... the length of an interval, i.e. the number of successive words between the repetititions of a particular word,
p_v .... the probability of an interval of length v,
P_v .... the cumulative probability, that is the probability of an interval length equal to v or less.

We consider an occurrence of a particular word whose relative frequency or probability is p. The probability of occurrence of any other word in the text is then $1 - p$. The probability that two neighbouring sites will be occupied by the word in question and another word is equal to $p(1 - p)$. The probability that of three successive sites the next two successive sites will be occupied by other words is $p(1 - p)^2$, and the probability of v other words between two repetitions of the word in question is then given by

$$p_v = p(1 - p)^v \tag{26}$$

We see from this that the probability of intervals of length v has its highest value for $v = 0$ since the bracketed expression to the 0-th power equals unity, and steadily decreases as v increases. This means that if the occurrences of the word in question are distributed by chance we would expect fewer gaps of length 1 than of length 0, fewer gaps of length 2 than of length 1, etc. The cumulative probability P_v is found by the addition of the values of p_v for $v = 0$, $v = 1$, $v = 2$, etc.

It has been shown (Vol. 1, Sect. 6.7) that for grammar words, for which intervals

[1] By Spang-Hanssen (1956). It is, however, to be regarded as only a rough approximation because of certain theoretical imperfections which cannot but affect its use.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

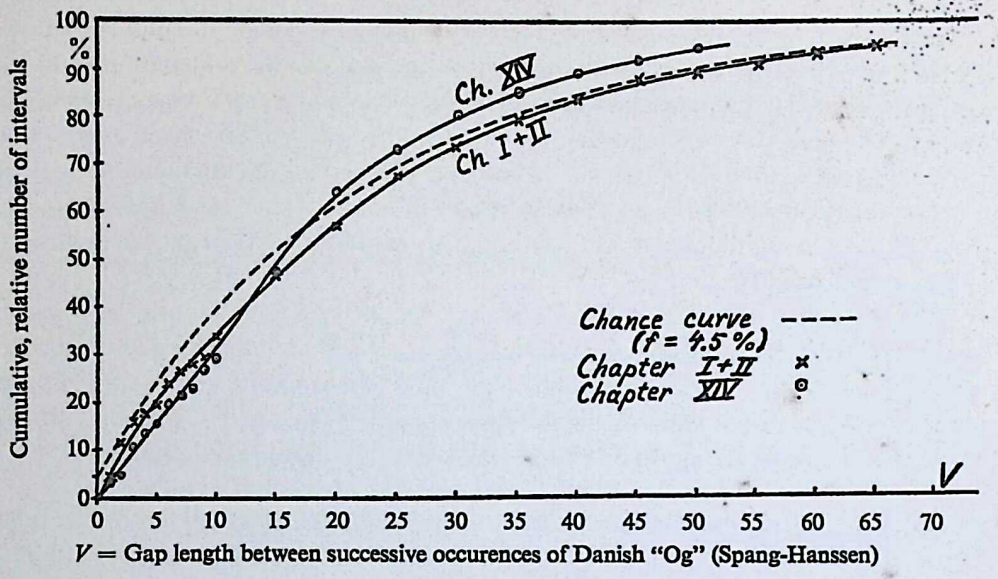


FIG. 12

of great length are rare, the particular chance distribution which fits the data is the Poisson Law. The advantage of using the Poisson law for the purpose of fitting a series of observed data is that numerical tables are available from which to read the values of p_v and P_v for a given average interval length, i.e. for a given v. This saves the trouble of having to calculate these values every time.

Illustration. The intervals between the repetition of the Danish word "*og*" meaning "*and*" were analysed in two separate parts of a modern Danish novel chosen at random, *Sonja Hauberg, Syv aar for Lea.* The first part, Chapter I and II comprises 130 gaps, totalling 3075 running words, thus the relative frequency of *og* is $p = 130/3075 = 4.2\%$. The second part, Chapter XIV, comprises 148 gaps, totalling 3198 running words, and $p = 148/3198 = 4.6\%$. Table 49 gives the observed cumulative relative frequency distribution of intervals of different length and the corresponding cumulative chance distribution (P). This distribution is fitted to the empirical values by putting the first term p_0 equal to 4.5% (Fig. 12).

Inspection of the table and of Fig. 12 shows a satisfactory agreement between the observed distribution and the chance distribution, except for small values of v.

11.2 THE REPEAT RATE

The repeat rate of a linguistic form, say a particular phoneme, is the probability that the two phonemes chosen at random from a great mass of phonemes representing a natural section of "la parole" of a language will turn out to be the same phoneme.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The mathematical expression for this is quite simple and straightforward. If the probability of a particular phoneme is p_t, the repeat rate is p_t^2. Summing over all the repeat rates for the different phonemes of the language in question, say n in number, we have

$$\text{Repeat rate} = \sum_{1}^{n} p_t^2$$

It is the probability that two phonemes picked at random from a great mass of phonemes will be repetitions of one or the other phoneme.

The relation between the gap distribution and the repeat rate is as follows. Provided gap length is expressed in the units whose repetition along the linear sequence of linguistic forms we are investigating, the squared reciprocal of the relative frequency as an estimate of the probability of the particular linguistic form, which is also the average gap length, is the reciprocal repeat rate of that particular form. If f is the

TABLE 49

Intervals between repetitions of Danish "og" (Spang-Hanssen)

Cum. Rel. Frequency of Intervals as observed			Cum. Probability of Intervals according to chance
v words	Ch. I+II %	Ch. XIV %	P %
0	0.0	0.0	4,5
1	3.8	3.4	8.8
2	11.5	4.7	12.9
3	16.1	10.6	16.8
4	17.7	13.5	20.6
5	19.2	15.5	24.1
6	23.8	19.6	27.5
7	26.1	21.6	30.8
8	27.7	23.0	33.9
9	29.2	27.0	36.9
10	33.8	29.1	39.7
15	46.1	46.6	52.1
20	56.9	64.2	61.8
25	57.7	73.0	69.8
30	73.8	80.4	76.0
35	79.2	85.1	80.9
40	83.8	89.2	84.9
45	87.7	91.9	88.0
50	89.2	94.6	90.5
55	90.7	.	92.4
60	93.0	.	94.0
65	94.6	.	95.2
70	.	.	.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

number of occurrences of a particular linguistic form, n the total number of such forms, $\bar{v}$ the average gap length, and p the probability of the particular form, we have

$$1/p^2 = (n/f)^2 = \bar{v}^2 \tag{27}$$

The distribution of gaps according to length plays no part in the structure of the repeat rate which, as seen, implies only the reciprocal of the *average* gap length. But the repeat rate parameter, although poorer in this respect, is more powerful in another, namely insofar as it admits the combination of the particular repeat rates into one overall characteristic of the linguistic text or of the language. The summation of all the particular repeat rates for the linguistic forms of a given level, Σp_i^2, gives an overall measure of repetitiveness as regards such forms.

The repeat rate is a characteristic of great usefulness for linguistic universes, not only in its application to the various levels of linguistic forms within a language, but also in its application to the comparison of different languages, and better than any other statistical parameter demonstrates the uniformity of language structure.

Another advantage is that to the linguist it is undoubtedly the summary expression of a well-known property of language as a system of communication, while its mathematical structure is extremely simple.

11.3 STATISTICAL ANALYSIS OF LANGUAGE CONTACT—THE REPEAT RATE OF INITIALS

11.3.1 Language mixture

Let us suppose that the comparison of style in two English writers has led to the conclusion that they differed with regard to the use of the Teutonic and Romance components of vocabulary. If the linguist has arrived at such a tentative conclusion and wishes to verify it, there seems to be no other way but to ascertain the etymological origin—preferably by using the Oxford Dictionary—of the words comprised in the vocabulary used by the two writers.

It needs no emphasising that the work required for the verification of the assumption is very considerable. However, much of this work could be saved if use is made of a rather surprising statistical relationship between the etymological provenancy of words and the alphabetical distribution of the vocabulary. Now, drawing up the frequency distribution of vocabulary according to the initials of words is just the type of work which, I am sure, would seem to most linguists and philologists nothing but a statistical exercise, and to most statisticians it would seem to be not even deserving of that name. The more surprising is it that it should lead, if done judiciously, to the conclusion that *if two English writers differed significantly in that respect, it was due to their vocabularies differing with regard to the proportions of the Teutonic and Romance components.* The information required for ascertaining such differences in the alphabetic distribution of vocabulary can be obtained from text samples of a reasonable size, and would be, if suitable electronic machinery is available, the work of a few minutes.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

As material we shall use vocabulary counts made by Yule on works of Bunyan and Macaulay. The following table shows the alphabetic distribution of vocabulary items—nouns only—for both writers.[2] For the purpose of comparison the numbers are also expressed per 10,000 nouns for each writer.

TABLE 50

Alphabetical distribution of nouns in the vocabularies of samples from Bunyan and Macaulay.[3] *Cols. 2 and 3 give the actual numbers of nouns with each initial letter, cols. 4 and 5 the distribution per 10,000 nouns*

1	2	3	4	5
Initial letter of noun	Number of nouns		Per 10,000	
	Bunyan	Macaulay	Bunyan	Macaulay
A	111	249	494	703
B	147	169	654	477
C	210	391	935	1104
D	153	237	681	669
E	69	162	307	457
F	112	150	499	423
G	72	107	321	302
H	110	112	490	316
I	72	172	321	485
J	22	27	98	76
K	18	22	80	62
L	84	122	374	344
M	124	209	552	590
N	40	52	178	147
O	41	73	183	206
P	188	338	837	954
Q	7	14	31	40
R	133	191	592	539
S	256	380	1140	1073
T	112	179	499	505
U	16	22	71	62
V	43	64	191	181
W	100	89	445	251
X	—	—	—	—
Y	5	8	22	23
Z	1	4	4	11
Total	2246	3543	9999	10,000

There seems to be a considerable difference in the use of words with a particular initial letter; for instance, Macaulay prefers words with initial A, C, E, whereas Bunyan prefers words with initial B, H, K, W.

[2] In samples of approximately 16,000 words each.
[3] Of approximately 16,000 words each.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Considering that English has two components, Germanic and Romance, which might conceivably differ in their alphabetic distribution, and further that the proportion of Romance vocabulary in Macaulay exceeds that in Bunyan considerably, it is a fair assumption to ascribe the difference in the alphabetic distribution between the two writers to the different proportions of Germanic and Romance vocabulary. Table 51 shows the number of nouns according to the two sources, Germanic and Romance, in each writer, and it is clear from Table 52 that the alphabetic distributions for the two components of English vocabulary are characteristically different from one another.

The question is now whether that discrepancy in the alphabetical distribution

TABLE 51

Etymological analysis of the alphabetical distribution leaving out words not classed (57 from Bunyan and 106 from Macaulay)

Initial letter of noun	Bunyan		Macaulay	
	Number of nouns			
	I. OE.-Teutonic	II. Latin-Romance	I. OE.-Teutonic	II. Latin-Romance
A	19	92	13	232
B	105	34	89	67
C	52	156	51	326
D	49	102	42	193
E	19	49	20	142
F	62	47	53	92
G	40	28	39	61
H	76	28	63	44
I	6	65	12	157
J	2	19	—	23
K	18	—	19	—
L	56	27	56	61
M	50	73	52	153
N	22	16	18	32
O	12	28	18	52
P	33	150	34	295
Q	1	6	4	10
R	36	94	34	154
S	141	109	141	231
T	54	55	62	110
U	8	4	7	14
V	2	40	1	62
W	94	4	79	7
X	—	—	—	—
Y	5	—	8	—
Z	—	1	1	3
Total	962	1227	916	2521

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 52

Bunyan and Macaulay; alphabetical distributions per 10,000 for the two etymological groups

1	2	3	4	5
Initial letter of noun	I. OE-Teutonic		II. Latin-Romance	
	Bunyan	Macaulay	Bunyan	Macaulay
A	198	142	750	920
B	1091	972	277	266
C	541	557	1271	1293
D	509	459	831	766
E	198	218	399	563
F	644	579	383	365
G	416	426	228	242
H	790	688	228	175
I	62	131	530	623
J	21	—	155	91
K	187	207	—	—
L	582	611	220	242
M	520	568	595	607
N	229	197	130	127
O	125	197	228	206
P	343	371	1222	1170
Q	10	44	49	40
R	374	371	766	611
S	1466	1539	888	916
T	561	677	448	436
U	83	76	33	56
V	21	11	326	246
W	977	862	33	28
X	—	—	—	—
Y	52	87	—	—
Z	—	11	8	12
Total	10000	10001	9998	10001

according to the source can be made to account for the observed difference between our two writers in that respect.

The analysis carried out by Yule was by means of the conventional statistical method of *correlation analysis*. We shall, however, make use of the *repeat rate* as a more appropriate parameter of statistical linguistics.

As a first step we transform the absolute numbers of Table 51 into relative frequencies by dividing them by the total of words in each class, Germanic or Romance. Table 53 shows the relative frequencies for words from either of these two sources side by side for both writers.

Inspection of this table shows clearly the close agreement between the relative

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 53

Relative frequency distribution of nouns according to the initial letter in the vocabularies of samples from Bunyan and Macaulay

Initial letter of noun	Germanic		Romance	
	Bunyan	Macaulay	Bunyan	Macaulay
A	.020	.014	.075	.092
B	.109	.097	.028	.027
C	.054	.056	.127	.129
D	.051	.046	.083	.077
E	.020	.022	.039	.056
F	.064	.058	.038	.037
G	.042	.043	.023	.024
H	.079	.069	.023	.018
I	.006	.013	.053	.062
J	.002	—	.015	.009
K	.019	.021	—	—
L	.058	.061	.022	.024
M	.052	.057	.059	.061
N	.023	.020	.013	.013
O	.013	.020	.023	.021
P	.034	.037	.122	.117
Q	.001	.004	.005	.004
R	.037	.037	.077	.061
S	.146	.154	.088	.092
T	.056	.068	.044	.044
U	.008	.007	.003	.006
V	.002	.001	.033	.024
W	.098	.086	.003	.003
X	—	—	—	—
Y	.005	.009	—	—
Z	—	.001	.001	.001

frequencies for the two writers for words from the same source. To establish the identity, or otherwise, of the series statistically, we have recourse to the repeat rate as a summarising parameter. It is obvious that the summation of the relative frequencies as they stand would give no answer to our question, since the series of probabilities for each writer must necessarily sum up to unity. On the other hand, squaring each relative frequency and summing will give a representative parameter for each series. The following table lists the repeat rate calculated by formula (27) Part I as the sum of the squared relative frequencies minus the reciprocal of the total:

	Bunyan	Macaulay	Difference
Germanic	.0731	.0717	.0014
Romance	.0724	.0739	.0015

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The standard deviation of the repeat rate, or rather of its estimate in terms of relative frequencies instead of probabilities, is calculated by (28) Part I as

$$\sigma_K^2 = \frac{0.0168\,N + 0.1182}{N(N-1)}$$

and results for the Germanic series as .004, which means that the fiducial limits within which the true repeat rate lies would be $\pm$ .012. The observed difference of .0014 does not even reach once the standard deviation and we may, therefore, regard the alphabetic distribution in Bunyan and Macaulay as random samples from one and the same distribution in the language. Similarly for the Romance words, for which $\sigma_x \sim .003$. Not only is the repeat rate for a given component sensibly the same for the two writers, but there is also no significant difference between the repeat rates for the different components in the same writer, which is rather remarkable.

This last fact should make us realise that the virtual identity of the repeat rate does not necessarily mean that the frequency distributions are the same: the essentially different distributions for the Germanic and Romance components also give the same repeat rate, and so will any mere permutation of the given frequencies or probabilities among the alphabetic symbols. This feature of the repeat rate is fully discussed in Section 5.2. In order to be able to say that a given component of the vocabulary, Germanic or Romance, had virtually the same frequency distribution in both writers, we must make sure that there was agreement between the probabilities attached to particular letters in both distributions. In other words, that the two distributions were truly random samples from one and the same population.

It can be shown mathematically[4] that in this case the quantity

$$\sum \frac{n_{i_1}\, n_{i_2}}{N_1\, N_2}$$

where n_{i_1} is the number of words with initial letter i in the sample of size N_1, and n_{i_2} the number of words with initial letter i in the sample of size N_2, was sensibly equal to the repeat rate of either of the two distributions. A significant difference between this quantity and $\Sigma(n_{i_1}/N_1)^2$ or $\Sigma(n_{i_2}/N_2)$ points to a real difference between the alphabetic distributions in the two samples. The quantity $\Sigma \frac{n_{i_1}\, n_{i_2}}{N_1\, N_2}$ calculated for the Germanic component results as .0728, and for the Romance component as .0726. We conclude, therefore, that the alphabetic distribution of vocabulary for each, the Teutonic or the Romance component, was virtually the same in the two writers. This being so, the difference in the alphabetic distribution of vocabulary noted at the beginning of this section, Table 50, can only be due to the components being present in different proportions of the vocabulary of each writer. This concludes the proof for the statement at the beginning of the section, according to

[4] Appendix C.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

which a *difference in the alphabetic distribution of vocabulary in two English writers admits the inference of their differential use of the Germanic and Romance components.*

In accordance with the inference drawn from the difference in the vocabulary distribution of the total vocabularies, we find the difference between Bunyan's and Macaulay's style in terms of the proportions which the two streams of vocabulary occupy quite pronounced:

	Bunyan	Macaulay
Teutonic	44%	27%
Romance	56%	73%

Regarding the general linguistic aspect of language contact, we have learnt from the above analysis that the relative frequency distributions of nouns in literary texts of English according to the initial letter for each, the Teutonic and the Romance component of vocabulary, can be regarded as random samples of the corresponding probability distribution in the language. As random samples, they are independent of text length and content; moreover, the distribution of each component, Germanic and Romance, is independent of, and without influence upon, the other. It is these findings which provide the theoretical basis for drawing the inference of different proportions of the Romance and Germanic components in the vocabularies of two English writers from an observed difference in the alphabetic distribution of their vocabularies.

Regarding the stylostatistical aspect, it is worth observing that the characteristic of the alphabetic distribution of words is, by no means, a trivial feature for distinguishing between individual style. After all, we use in civil life the family name plus initials for identifying people, the underlying idea being that it is very unlikely for a given combination of these symbols to repeat itself in the community. If it were not so, identification by means of initials and surname would be illusory. Similarly in language, it is very unlikely that the style of different individuals should be the same in terms of the distribution of initials.

11.3.2 Loanwords

If we think of contact between languages in terms of loanwords, it is of interest to enquire whether we can speak of certain laws governing the entry of words from one language into the other. As will be seen, the alphabetical distribution of loanwords enables us to establish a certain statistical law of that type.

Guiraud (1958)[5] has used the alphabetical distribution—more correctly the distribution according to the initial phoneme—of loanwords in French for establishing what he calls the "équilibre phonologique", by which he means that it is the equilibrium

[5] P. Guiraud, "Emprunts et équilibre phonologique", *Zeitschrift für Romanische Philologie*, 74 (Tübingen, 1958), pp. 78-88.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

in the use of phonemes in French which governs the entry of words from other languages. His method was as follows: using the *Dictionnaire Générale*, he recorded the number of French words of foreign origin according to the original language: German, Dutch, English, Provençal, Italian, Spanish etc., and their alphabetical distribution according to 26 initials. He then summed the words with the same initials from the six countries and got the overall alphabetical distribution of loanwords in French. This distribution he compared with what could be expected if the loanwords from a given country represented a random sample of the alphabetical (phoneme) distribution of all the words in the dictionary of that country. He again summed these expected frequencies for all the countries and got the theoretical distribution of loanwords in French under the assumption of random sampling. The comparison of the two series then made him conclude that although for some letters the actual number of loanwords was in agreement with what could be expected on random sampling, for certain initials this was not the case. This led to the further conclusion that it was a deficiency in the French phoneme distribution which resulted in the acquisition of loanwords in excess of what could be expected by random sampling.

The method as employed by Guiraud is formally not correct.

1. He uses the square root of the expected number of loanwords summed over all the countries in question as the standard deviation of a Poisson distribution, which implies that the probability of occurrence of words with a given initial was the same for every language taken into consideration. This is evidently not the case, and in a way contradicts the very idea of the experiment which requires the estimation of the particular probabilities for words with a given initial in each of these languages.

2. The concept of a phonological equilibrium of French as the governing factor for the entry of loanwords, makes one think of the total phoneme distribution in the language, that is of the solidarity between all phonemes, and not only of that between initial phonemes.

3. But even if we were to restrict the idea of the phonological equilibrium to the distribution of initial phonemes, there is no explanation how the deficiency in words with a certain initial would lead to an excess in borrowing of words with that initial. Without an acceptable hypothesis of the mechanism that would produce such an excess entry of words, the argument appears to be of the post hoc ergo propter hoc-type.

A scientifically more correct procedure consists in comparing the alphabetical distribution of loanwords from a particular language with the alphabetical distribution from the dictionary of that language. This enables one to say for each language separately whether the observed alphabetical distribution of actual loanwords could be regarded as a random sample of that in the dictionary. For the purpose of the comparison, the data used are from Guiraud's tables. The last three initials of Guiraud's table of observed loanword numbers, Y, Z, W, comprising very few words were left out because, although he had given the corresponding frequencies of words

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 54

Alphabetical distribution of loanwords in French (after Guiraud)

	German	Dutch	English	Provençal	Italian	Spanish
A	2	3	13	26	51	27
B	24	24	23	76	122	34
KA	5	9	6	63	112	39
K	14	6	27	20	52	20
CH	3	—	3	3	10	10
D	5	11	14	14	39	12
E	8	8	7	27	69	7
F	6	5	13	19	60	3
GA	2	—	3	15	18	2
G	5	4	6	10	29	12
J	1	—	5	7	10	15
H	11	5	9	1	5	8
I	—	—	3	1	22	3
L	3	6	9	8	22	9
M	1	10	9	24	64	26
N	1	2	—	2	11	5
O	1	1	—	1	10	6
P	2	3	14	18	74	14
R	3	—	10	10	37	14
S	8	5	23	13	74	24
T	5	2	18	8	51	33
U	—	—	—	—	—	4
V	8	6	3	4	25	7
Total	118	110	218	370	967	278

with these initials separately in the list of observed loanwords, he has pooled the figures for X, Y, Z in his list of expected frequencies and given no figure for W.

As a suitable statistical parameter we use the repeat rate of loanwords. The following table lists the repeat rate for the observed alphabetical distribution of loanwords calculated according to formula (27) Part I as the ratio of the sum of the squared frequencies of Table 54 divided by the square of the total minus the reciprocal of the total. The repeat rate of the dictionary percentages is calculated by formula (25) as the sum of the squared probabilities, because the dictionary values can be taken—at least they are meant to be taken—as the true probabilities in the language. Each repeat rate of the series of the alphabetical distribution of loanwords is given with its standard deviation calculated by formula (28). The column "Fiducial Limits" gives the range as the repeat rate $\pm 3\sigma$ within which the true repeat rate will lie with a very high probability.

The results are most interesting. The dictionary repeat rate for all languages except Provençal lies within the fiducial limits of the loanword repeat rate which means that for all these languages we may regard the alphabetical distribution of loanwords in French as a random sample of the alphabetical distribution in the dictionary of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 55

Alphabetical distribution in the dictionary (%) (after Guiraud)

	German	English	Dutch	Provençal	Italian	Spanish
A	8.1	10	5.6	5.8	7.7	13.0
B	6.1	7.2	7.3	6.6	7	4.9
KA	0.7	2.7	2.1	4.7	5	5.1
K	4.4	6.3	7.2	3.1	5.4	8.2
CH	—	1.2	—	1.7	0.9	1.4
D	3.1	5.7	4.1	6.0	3.3	5.6
E	5.02	4.0	2.1	10.1	6	7.4
F	5.5	5.0	1.6	5.4	5.6	2.8
GA	0.3	0.8	0.6	1.6	1.4	0.9
G	6.5	2.4	5.3	2.3	3.6	1.7
J	0.7	1.6	0.9	1.3	1.2	1.1
H	5.6	4.0	4.1	1.1	0.1	2.6
I	0.6	4.5	2.2	0.5	2.2	2.5
L	3.6	2.9	3.7	3.1	3	3.2
M	4.5	4.8	4.2	8	7.5	6.2
N	3.0	1.5	1.9	1.2	1.4	1.5
O	0.7	1.9	5	1.5	1.4	1.7
P	2.0	7.8	6.5	11.0	8.7	8.5
R	3.2	4.2	4.3	8	4.3	4.9
S	13.2	11.3	11.1	5.4	12.5	7.0
T	3.2	4.8	4.8	6.2	5.6	5.5
U	4.2	0.6	1.5	0.2	0.7	0.6
V	6.1	1.4	7.3	3.5	3	2.7

TABLE 56

Repeat rates

	Loanwords	Fiducial Limits	Dictionary
German	.0823 ± .012	.1183 .0463	.0559
Dutch	.0875 ± .013	.1265 .0485	.0538
English	.0708 ± .009	.0978 .0438	.0587
Prov.	.0982 ± .007	.1192 .0772	.0632
Italian	.0683 ± .004	.0803 .1563	.0622
Spanish	.0689 ± .008	.0929 .0449	.0637

respective language. For Provençal the theoretical repeat rate is slightly below that of the lower limit of the observed repeat rate, which must be interpreted as an interference with the randomness of the sample of loanwords. Such interference is not

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

surprising since the entry of loanwords is, no doubt to some extent, governed by semantic principles and conceivably also by the distribution of initial phonemes of the borrowing language, French in this case.

What is surprising is that for the five other languages we find such good agreement between what might be expected on pure chance. This does by no means exclude the conscious acquisition of particular loanwords by the speakers of French because of semantic needs for such words. Since the linguistic symbol, and more particularly the initial letter, is very largely independent of the content—as was stressed by de Saussure—, it follows that the loanwords chosen originally for semantic reasons may yet, with regard to certain properties of the alphabetic symbols, be random samples of that distribution in their own language. The reader is referred to the section on the ergodicity of linguistic masses (Sect. 20.3).

There is, however, a certain feature which makes it advisable to qualify somewhat the basic conclusion as to randomness of the loanword distribution according to initial phoneme. The repeat rates in the first column of Table 56 are throughout higher than the corresponding ones in the last, which means a slightly higher degree of repetition through certain initial phonemes occurring oftener than expected. This interference with the randomness in respect of one or the other initial phoneme, though not enough to reach significance for the whole series, may yet point to the distribution of initial phonemes in French as exercising some influence upon the selection of loanwords, apart from the fact that it may also be due, to some extent, to the semantic factor: borrowed words belonging to the same field may all have the same initial, which would tend to increase the number of such words beyond the expectation on pure chance.

The general conclusion reached with regard to loanwords is in good agreement with that reached in the previous section with regard to language mixtures. There we found that the text samples of different length, and different content, could be regarded as random samples of the corresponding distribution in the language, and that such distributions for etymologically different components of the language did not influence one another.

Here we have found that the batch of loanwords taken over from a foreign language retains by and large the alphabetical distribution in the dictionary of that language to such an extent that it can be regarded as a random sample of that distribution. This means that insofar the sample of loanwords is independent from the alphabetical distribution in the borrowing language, French in our case.

Considering what has been said above about the slight interference by selection due to a possible influence of the initial phoneme distribution in French, the difference between Guiraud's results and ours may be formulated like this: according to Guiraud the distribution of initial phonemes in the borrowing language is the major factor in the selection of loanwords from other languages; according to our results the main factor for the distribution of loanwords according to initial phonemes is the language of origin, compared with which the influence of the borrowing language is slight.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

THIRD PART

INFORMATION THEORY

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XII

THE CODING OF LANGUAGE IN THE BRAIN

12.1 THE LINGUISTIC INTERPRETATION OF INFORMATION, ENTROPY, REDUNDANCY[1]

Information theory owes its origin to the discovery that relative expenditure in terms of the relative frequencies of the ultimate linguistic units, such as phonemes or letters, is independent of the content of the message. This enables one to measure the "cost" of transmitting any message, that is messages in general. If we speak of that measure as one of information, it refers primarily to the cost of transmitting information.

Apart from this, the implied stability of relative frequencies of linguistic symbols enables us to make relevant guesses at missing parts of the message, and thus obtain "information" in a formal sense, that is information about the linguistic expression of the content, and about the efficiency of the linguistic code used.

It is paradoxical, but nevertheless true that perhaps in no other branch of knowledge is there so little communication (of thought) between those interested in it, as there is in communication theory. To mention only one thing: it is today pretty generally acknowledged that what Shannon has called information is linguistically the lack of information. This has necessitated the introduction of the negative entropy—the negentropy—and the use of the concept of Ektropy when applying information-theoretical ideas to language (Vol. I, Sect. 10.8).

A small entropy H means that the amount of information which is attached to a symbol is small. This has the consequence that a missing symbol, or one whose linguistic value is not known, does not mean a great loss to the understanding, and since the probability of such a symbol is great—otherwise $\log 1/p$ would not be small—, a guess can be made with good expectation of being correct. Success in guessing depends upon great probability, and thus also upon small H. It follows that the smaller the "information" attached to a symbol, the greater the chances of restoring or deciphering the message. A small H thus means a small unexpectedness, a large H a great amount of unexpectedness. This leads to the conclusion that "information" in terms of H does not mean what it says, but rather the opposite, namely the lack of information.

There is, however, an explanation of why Shannon and his fellow engineers speak of H as "information". What we have in mind in the above interpretation is what may

[1] The reader not sufficiently acquainted with information theory is advised to read Sects. 12.2 to 12.4 before Sect. 12.1.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

be called "advance or past information" embodied in the structure of the language and immediately accessible to all who know the language; what the communication engineer has in mind is "future information". In the first case we ask: "How much information *have I got* that will enable me to restore a missing linguistic form?" In the second: "How much information *shall I get* by restoring the missing linguistic form?" The answer to the first question is "$-H$"; to the second, "$+H$".[2]

It is quite similar with the term "Redundancy", as used by Shannon. According to Brillouin, it means correlation between unit symbols, say letters of printed English. Differently expressed, it implies certain constraints to which the sequence of letters is subject, and which make it differ from a random sequence. But this is again important "advance information" for the man who knows the language, which helps him to restore missing parts and to make successful guesses. To him it is thus anything but "redundant"! The term is thus linguistically quite wrong. It means again the opposite of what it says, just like the term "information". It should be replaced by "structure" or "constraints" and "information" by "unexpectedness".

In particular, it is the conventional interpretation of Redundancy which appears to be the stumbling block for a sensible application of information theory to language structure. Since order, in the sense of a well-defined gradient of probabilities, must be regarded as a sign of language structure, it follows that a smaller entropy, H, will—for a given series of language forms—indicate a higher degree of language structure, and the entropy obtained from an equidistribution of the probabilities, H', be a sign of lack of structure, and thus of disorder or chaos. Further, since redundancy is defined as the complement to unity of the relative entropy H/H',

$$R = 1 - H/H',$$

it follows that it is greater, the smaller H, and thus, the greater the state of order or the degree of structure in the language.

That this interpretation of the redundancy, which is in complete accordance with the mathematical structure of information theory, should almost invariably meet with objections from information theorists, and, in particular, why a characteristic and valuable language property, which is quite essential for the quick and correct differentiation between its forms, should be called "redundancy", has puzzled me for a long time until I hit upon the true explanation.

Information theory was devised and first used for deciphering secret codes (Shannon, Weaver). A basic principle of secret codes is that they should *not* possess the statistical characteristics of language. A first approximation to this is that the relative frequencies of all n different signs of a cryptogram should be the same, $1/n$. Therefore, the cryptoanalyst starts his work on the cryptogram by examining the hypothesis that all cyphers have the same probability of use. Or, what is the same, under the hypothesis that the cryptogram has arisen through random sampling from a multinomial distribution with n attributes, all having the same probability $1/n$.

[2] For a numerical illustration, see Sect. 9.3.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Considering now that the ideal cryptogram is characterised by a rectangular distribution of frequencies, or by the complete lack of a gradient, it follows that the entropy H calculated from it will closely approach to the maximum entropy H' for the given degrees of freedom, and the redundancy R will approximate to Zero. The absence of linguistic structure, which is a disadvantage for normal languages, is a virtue for the secret code, and the cryptoanalysts naturally used for the lack of that property a name indicating something undesirable. And since what is undesirable here is that some signs should be more often repeated than others—which would make it more similar to a natural language—, they called it "redundancy". Had they thought of natural languages instead of secret codes, they would have chosen another name, more in accordance with the fact that a well-defined gradient of probabilities was a most desirable and necessary feature of language.

It will be shown that it is not only the interpretation of the term "information" which requires revision when applied to linguistic problems, but equally well its mathematical formulation as the logarithm of a probability, the binary units in which information is measured, and also the problem whether to calculate the entropy per word or per letter or syllable contained in a word (Chapters XII, XIV B, XIII respectively).

12.2 THE UNITS OF INFORMATION

Our brain and nervous system work in such a way that a nerve cell is either excited or not, which means that it can assume either of two mutually exclusive states, but not both at the same time. This is in accordance with the principle known in logic as that of Contradiction, and is implied in what is called a Yes-No decision, all or nothing decision, or a 0-1 alternative.

In symbolic logic, as set forth by G. Boole in his *Laws of Thought* (1854), the principle of contradiction is given the algebraic form

$$x(1 - x) = 0$$
$$\therefore \quad x = x^2$$

and the name of the law of Duality. It proves to be indispensible for the system of symbolic logic, that is for the algebraic presentation of the laws of logic, and, since language is, in the last resort, only the formulation of logical relations in terms of linguistic forms, it seems sensible to conclude that a dual or dyadic symbol system will be appropriate for language.

The algebra which we use for the development of symbolic logic—known as Boolean algebra—is one in which the symbols x, y, z etc. admit only of the values 0 and 1. Similarly, the formula for the entropy which, according to its definition, is calculated as the average of the logarithm of the reciprocals of probabilities of the different states of a system, has for language a meaning only if the logarithm is the dyadic one, implying that information is obtained from a succession of suitable alternatives, and thus by means of 0 or 1 decisions.[3]

[3] See Epilogue, Sect. I.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

To the Yes-No principle there corresponds a coding system with only two symbols. An application is the Morse code of telegraphy with its Point-Dash symbol system. If for a number system of this kind one chooses 0 and 1 as symbols, one has the so-called Dual, Dyadic or Binary number system in which all numbers are written by means of 0 and 1 only (Table 57).

TABLE 57

Decimal System	Binary System
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111
16	10000

Whereas in the decimal notation the symbols are the coefficients of powers of ten, in the binary notation they are the coefficients of powers of two. E.g.

$$10 = 1.2^3 + 0.2^2 + 1.2^1 + 0.2^0$$

which is written in binary notation as 1010.

12.3 TRANSMISSION OF INFORMATION USING A BINARY CODE

The following derivation of the basic principles of Information theory is based upon the exposition of the theory by the Russian mathematician A.N. Kolmogoroff, submitted to the Academy of Science of the U.S.S.R. in 1956. It is of special interest that Kolmogoroff—like the present author in Vol. I, Chapt. 10—regards the binary symbolism as an essential feature of the theory when applied to language, a view not shared by some scientists (Epilogue, Sect. I).

Turning to the problem of the transmission of information in the form of communication by telegraph or similar apparatus, we start with the very simple problem of transmitting a sequence of decimal numbers

$$a_1, a_2, \ldots\ldots a_k, \ldots \qquad (1)$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

by means of a sequence of binary numbers

$$b_1, b_2, \ldots\ldots b_k, \ldots \tag{2}$$

Let us assume that sequence (1) arises in unit time per single symbol. We then ask with what speed, or at what rate, the binary symbols are to be transmitted, if it is required that the reception in terms of sequence (2) of the message sent originally in terms of sequence (1) is to be accomplished without loss of time.

For transmitting all single signs of (1), that is numbers up to 10, by binary signs, four such signs are required (since $2^3 = 8 < 10$ and only $2^4 = 16 > 10$). For transmitting all pairs of decimal symbols, that is the numbers from 11 to 100, by groups of binary units, seven such units are required (since $2^6 = 64 < 100$ and only $2^7 = 128 > 100$). For transmitting all triples of the symbols of sequence (1) in binary units, ten such units will be sufficient (since $2^9 = 512 < 1000$ but $2^{10} = 1024 > 1000$), and so on. In words: the number of binary symbols required for transmitting a decimal number is its dyadic logarithm rounded off to the next whole number.

The speed of formation of the signs of sequence (2) corresponding to this system of transmission, i.e. the number of binary signs which are required, at an average, for transmission per unit time equals the ratio of the logarithm of a number in the binary and decimal systems rounded off to the next whole number:

$$v_1 = 4; \quad v_2 = 7/2 = 3.5; \quad v_3 = 10/3 = 3.33 \ldots$$

It is easily proved that the series of numbers arising in this way converges and approaches to the limit

$$v = \log_2 10 = 3.322 \ldots$$

This means that v represents a lower limit for the speed or rate of transmission of the symbols of sequence (2), if by means of it those of sequence (1) are to be transmitted without a systematic loss of time.

Already in this purposely simple illustration we may trace some characteristic features of communication which have come to light through the modern theory of information.

1. In spite of *qualitative* heterogeneity of messages, that is their differences as regards content, it is possible, and reasonable, to attempt to measure *quantitatively* the rate or speed of transmission.

For the decimal series, the speed of transmission if binary symbols are used is, on an average,

$$I = \log_2 10 = 3.322 \ldots \tag{3}$$

This means that per unit time, or for every decimal unit, we require 3.322 binary units.

2. Let us suppose we were engaged in the well known game of guessing an object by asking questions which may only be answered by "yes" and "no". To fix our ideas, we will assume that a person is required to guess a letter in an alphabet of 32 letters. If that person has a mathematical mind, he or she would ask whether the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

letter was in the upper half of the 32 letters,[4] then whether it was in the upper half of the 16 letters in which the first question had located it, and so forth. By five such questions we can determine the letter. Writing now "1" for "yes" and "0" for "no", the letter could be expressed by a symbol such as 01010, where the first digit is the answer to the first question, the second that to the second question, and so on.

This suggests that the number of digits, that is the number of "yeses" and "noes", by which the letter is fixed, be taken as a measure of information. It measures not really the amount of information—or the content—, but the action needed for obtaining it. Thus we can say that for an alphabet of 32 letters, the amount of uncertainty attached to a letter is never more than 5, since with 5 guesses, at the most, we can ascertain the letter somebody had in mind, or a missing letter. But 5 is the logarithm of 32 to the base 2, or as we say, its dyadic logarithm. In general, the information required for a correct guess or identification of a particular item out of N equally probable items is

$$I = \log_2 N, \tag{4}$$

and since $\log_2 = 3.322 \log_{10}$,

$$I = 3.322 \log_{10} N.$$

12.4 ENTROPY AS A MEASURE OF INFORMATION

We now turn our attention to a particular aspect of literary texts, viz., the different frequencies with which the different letters of the alphabet occur in them. If we denote by the numbers

$$n_1, n_2, \ldots\ldots n_k$$

the frequencies with which in a text of length $n = n_1 + n_2 + \ldots\ldots + n_k$ each of the letters

$$a_1, a_2, \ldots\ldots a_k$$

occur, then the number of literary texts of length n composed of these letters with the frequencies assigned to them results by combinatorial mathematics as

$$N = \frac{n!}{n_1!\, n_2! \ldots n_k!} \tag{5}$$

Using Stirling's formula for the computation of factorial numbers which in its approximate form sufficient for our purpose is

$$\log(n!) \sim n \log n,$$

it can easily be shown[5] that for great n_i

$$\log N \sim -n \Sigma p_i \log p_i \tag{6}$$

[4] In the order of the alphabet.
[5] Vol. I, p. 165.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

where $p_i = n_i/n$, the proportion or probability of occurrence of the individual letters. (Here, and throughout this section, we denote by "log" the log to the base 2, even if this is not expressly stated.)

Since, as we know from (4), the logarithm of a number can be regarded as a quantitative measure of information, we write, using the symbol I for "information"

$$I = -n \Sigma p_i \log p_i \tag{7}$$

This result may be expressed as follows: using the letters of the alphabet independently with probabilities p_i, the quantity of transmitted information per letter in the text equals

$$I = -\Sigma p_i \log p_i \tag{8}$$

For the case of equiprobability of letters

$$p_1 = p_2 = \ldots\ldots = p_k = 1/k$$

we obtain

$$I' = \log k \tag{9}$$

In general, for arbitrary differences between the p_i,

$$I < I' \tag{10}$$

In the following illustration we consider the transmission of a text in Russian. Since there are 33 different letters in the Russian alphabet, the number of different texts of length n composed of these letters is

$$N = 33^n.$$

The quantity of information required for correctly guessing a particular one of these texts is, according to (4)

$$I = \log 33^n = n \log 33$$

Thus, the information required for correctly guessing a particular text, is the logarithm of the possible number of texts of such length.

And the reason for calling the dyadic logarithm of the possible number of texts of length n "information" is that it represents the number of alternatives (binary choices) required for guessing or identifying a particular text.

12.5 A DIFFERENT DERIVATION OF THE ENTROPY AS A MEASURE OF INFORMATION

Let us consider a situation with N_0 different possible outcomes, which are equally probable a priori. This is the initial situation, when we have no special information I_0 about the system under consideration. If one of the N_0 outcomes is realised, this would mean that we have obtained more information I_1 about the system. The greater N_0, the larger will be the amount of information required for making the selection of a particular outcome. In symbols, we have:

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Initial situation: $I_0 = 0$ with N_0 equally probable outcomes.

Final situation: $I_1 \neq 0$ with $N_1 = 1$, i.e. one particular outcome. The symbol I denotes information, and the definition of information is

$$I_1 = K \ln N_0 \tag{11}$$

where K is a constant and ln means the natural logarithm (logarithm to the base e).

The use of a logarithm in formula (11) is sometimes justified by formal reasons, such as that we want information to be additive. For let us consider two independent problems, the first having N_{01} and the second N_{02} a priori equally probable solutions. Each solution of the first problem can be coupled with any solution of the second one. Hence the total number of such combinations is

$$N_0 = N_{01} \cdot N_{02} \tag{12}$$

and

$$I_1 = K \ln(N_{01} \cdot N_{02}) = K \ln N_{01} + K \ln N_{02} = I_{11} + I_{12} \tag{13}$$

The definition of information can be generalised to cover the case when N_o possibilities exist in the initial situation, while the final situation still contains $N_1 \neq 1$, i.e. more than one, possibilities:

Initially: $I_0 = 0$ with N_0 equally probable cases

Finally : $I_1 \neq 0$ with N_1 equally probable cases.

According to the definition of the Information, we have in such a case:

$$I_1 = K \ln (N_0/N_1) = K \ln N_0 - K \ln N_1 \tag{14}$$

This definition reduces to equ. (11) when $N_1 = 1$, and represents an obvious generalisation of (11).

Writing equation (14) as

$$I_1 = -K \ln (N_1/N_0)$$

and writing for $N_1/N_0 = p_1$, we have

$$I_1 = -K \ln p_1 \tag{15}$$

which defines information as the negative logarithm of a probability. As a rule, we work with a logarithm to the base 10, $\log_{10}$. By a suitable choice of the constant K the logarithm is transformed into one to the base required.

To say, however, that we use the logarithm of a probability for measuring information because this makes information additive is the reverse of the order in which the respective concepts have entered the mind of the scientist concerned with information theory. The reader of such an interpretation must get the impression that the scientist first decided that information was additive, whereupon he selected the entropy as a suitable measure of it. Apart from the fact that there is absolutely no a priori reason why information in the general sense of the term should be additive, the true sequence is the reverse of this. The entropy as a well-known measure of disorder was found to

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

be suitable for measuring information in the sense explained above, and since the entropy $I = -\Sigma p \log p$ appears as the expectation of $\log p$, that is algebraically as its arithmetic mean or as the weighted sum of $\log p$, it follows that information in terms of $\log p$ is additive. Briefly: *the entropy I is not used because information is additive, but information is additive because the entropy was chosen as an average measure of information.*

The reason for calling the dyadic logarithm of the possible number of events of a certain type "information" is that it represents the number of alternatives required for successfully identifying or guessing a particular event (Sect. 12.4).

12.6 THE STORAGE OF INFORMATION IN THE BRAIN

Since the area of the brain devoted to language is limited to one hemisphere only —the left—, and thus rather restricted in extension, the principle of efficient storage must be all important. As we shall show, such storage provides the systematic component of linguistic information which enables us to reduce the aleatoric (chance) component, measured by the entropy, to a minimum.

The linguistic code achieves such storage by three principles:

1. A distinct gradient of probabilities as against equidistribution of symbols,
2. A system of transition probabilities between the single symbols of the type of the Markoff chain (see Chpts. X, XIVB).
3. A lower limit to the product of the variabilities, of the type of the Uncertainty relation.

We shall demonstrate these conclusions by considering imparting information by means of the English alphabet. Point 3 will be dealt with in Sect. 13.3.

Let the a priori probability of the j-th letter in the alphabet plus word space or "blank" be p_j ($j = 1, 2, \ldots 27$), as specified in table 58, cols. 1, 2. In col. 3, the values of $\log_{10} p$ are listed. By making $K = 3.322$, they are transformed into the values of the dyadic logarithm or $\log_2 p$.

The computational work is facilitated by a numerical table (Appendix) for the values of $-3.322\, p_i \log_{10} p_i$ for all values from $p_i = .01$ to $p_i = .999$ (Previtera, 1958). The values in the body of this table are multiplied by 1000; all we have to do, therefore, in order to calculate the entropy for a given series of p_i, is to read the corresponding value in the body of this table, put a decimal point in front of it and sum over all the values of p_i.

For our language the complete alphabet contains 27 symbols: 26 letters plus the "blank" or space in between words. If the language would use all these symbols to the same extent, in other words, if all symbols were equally probable a priori, the information contained in a sentence of N letters would be[6]

$$H' = \log_2 27 \text{ bits}$$

[6] From now on we shall use H as the conventional symbol for the entropy.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 58

The probability of occurrence p and values of $-\log_{10} p$ for the letters of the English language (Brillouin 1956)

Symbol	Probability p	$-\log_{10} p$	Symbol	Probability p	$-\log_{10} p$
Word space, or "blank"	0.2	0.699	L	0.029	1.54
E	0.105	0.979	C	0.023	1.64
T	0.072	1.143	FU	0.0225	1.65
O	0.0654	1.184	M	0.021	1.68
A	0.063	1.2	P	0.0175	1.76
N	0.059	1.23	YW	0.012	1.92
I	0.055	1.26	G	0.011	1.96
R	0.054	1.27	B	0.0105	1.98
S	0.052	1.28	V	0.008	2.1
H	0.047	1.33	K	0.003	2.52
D	0.035	1.46	X	0.002	2.7
			JQZ	0.001	3.0

or $$H' = 4.76 \text{ bits per letter}$$

However, as letter counts show, different letters do not occur with the same probability; some letters like e, t, n are preferred in our language as compared with k, f, g. The actual probabilities of letters for the English language are given in Table 58.

Calculating the average information per letter according to formula 8, that is as

$$H = -K \sum_{j=1}^{27} p_j \log p_j \tag{15}$$

we find it as 4.03 bits/letter. This means in our terminology that the information of the system of letters is somewhat reduced, or that the information embodied in a system of letters can be stored in a smaller space by using special information about the different probability of occurrence of different letters.

However, it is well known that a language is characterised by preference and abhorrence of certain sound (phoneme) combinations. As far as phonemes are represented by letters, we may say that different letter clusters will, in general, have a different probability in a given language. The probabilities for letter digrams and trigrams have been established for the English language, though insufficient data exist on larger groups. If this additional information is duly taken into account when calculating the information inherently provided by the system of letters, we find the entropy reduced to $H = 3.32$ bits/letter when digrams are considered and $H = 3.1$ bits/letter when trigrams are considered. Shannon arrives at the limiting value of the probability if the transition probabilities in larger letter groups are duly taken into account, of 2.14 bits/letter and in another set of experiments of 1.9 bits/letter.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

This shows that the structural properties of written English are such as to make it possible to reduce the aleatoric component—on that level of language at least—to a minimum.

Strictly speaking, H does not represent the total of information, but what is required in addition to some stored knowledge. When letters are used freely with equal a priori probabilities, the information is 4.76 bits/letter. If we impose the constraints corresponding to Table 58, taking into account the different a priori probabilities as established in the language, the information per letter drops to 4.03 bits; and if the transition probabilities proper to the structure of the language are taken into account, the information drops still further to about 2 bits/letter. The constraints may be regarded as representing certain advance information about the letters to be selected, which enables us to reduce the information in terms of H calculated from the letters accordingly. We may therefore regard the information calculated when all the constraints in question are properly taken into account, as that part of information which is required in addition to the advance information embodied in the constraints. In this sense we may speak of the entropy as the new information required in addition to the stored information.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIII

THE MATHEMATICAL CONDITION FOR OPTIMAL REDUNDANCY

13.1 EFFICIENCY OF ARTIFICIAL CODES—PITMAN'S SHORTHAND AS A CASE OF OPTIMAL CODING

During examination of artificial codes such as Pitman's Shorthand, which were designed with a view to obtaining maximum efficiency and convenience,[1] it was observed that the frequencies of symbols with 1,2 ... N "strokes" (in the sense in which that term is defined below) seemed to approximate closely to the terms of a geometric series with common factor 0.5 and, consequently, a mean 2.

The principle of Pitman's shorthand is that a geometrical alternative or opposition is used to indicate a phonetic one. Thus, different directions and thickness of straight strokes are used to indicate different types of consonants; the alternative of direction for one and the same outline is used to indicate whether the consonant represented by the outline is followed, or is not followed by a vowel; the alternative between the right and left hook at the start of the outline is used to indicate whether *r* or *l* follows a consonant; the alternative between the right and left end hook to indicate whether a word closes with an *n* or an *f*; the alternative between the forms o and) to indicate whether a vowel precedes the letter *s*; the doubling principle to indicate the derivative affix *-er*; the halving principle to indicate the affix *-ed*; the large end hook for the syllable "*tion*" in two directions to indicate whether the preceding vowel is short or long; the device of a point put in front or at the end of the outline to indicate the prefix *con* or the ending *-ing*, respectively; the alternative between the three positions of the outline to indicate the type of the first vowel in the word.

[1] The statement about what constitutes efficiency in a linguistic code on p. 181, Vol. I, is a condensed version of the conclusion as reached in the author's paper of 1953 (*Metron*, Vol. 17, No. 1 & 2): "What can be inferred from all this is that redundancy, which depends merely upon the degree of compression of H against H', and thus upon the presence of a sort of gradient, determines the efficiency of coding not entirely by itself, but in conjunction with another element. This element is the property of the gradient to extend over a certain length, thus ensuring a fair range of variation among the symbols ..... Thus the efficiency of the linguistic code, whose symbols are words of different length in terms of syllables, appears to be the resultant vector of two component ones, viz., redundancy due to a gradient of probability so as to ensure sufficient differentiation for guessing word length, and length of gradient so as to ensure sufficient range of the variable word length. The result will be a closer approximation of H to M."

To put it briefly: we know that H should be smaller than H' if the code is to be efficient; but there must be a limit to the decrease of H relative to H' if the code is still to be of any practical use. In what follows, this limit will be established.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC:0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Every outline in a piece of writing was examined with regard to the number of such alternatives or principles it contained, without counting that of position of outline since it occurs in every word, and thus would merely raise the count by unity in every outline. The following table shows the distribution of outlines of words—contractions of words were not included—in two pieces of writing taken from a shorthand text of Goldsmith's *Vicar of Wakefield* (edited by J. Pitman, London) and the fiitted geometric series.

TABLE 59

Distribution of Length of Pitman Shorthand Outline

Number of alternatives ("strokes")	Vicar of Wakefield				Sample 1 & 2 combined		Geometric Series (c.f. = .5457) $M = 2.2$	Geometric Series (c.f. = .5) $M = 2$
	Sample I		Sample II					
	f	‰	f	‰	f	‰	‰	‰
1	503	478	445	431	948	454.3	454.3	500.0
2	260	247	264	256	524	251.2	247.9	250.0
3	143	136	136	132	279	133.7	135.3	125.0
4	100	94	108	104	208	99.3	73.8	62.5
5	41	39	56	54	97	46.2	40.3	31.3
6	6	6	19	18	25	12.0	22.0	15.6
7			5	5	5	2.4	12.0	7.8
							6.5	3.9
							3.6	2.0
							2.0	1.0
							1.1	0.5
							0.6	0.3
							0.3	0.2
							0.16	0.1
Total	1,053	1,000	1,033	1,000	2,086	1,000.0	1,000.0	1,000.0
Average number	1.99		2.17		2.08			
Entropy H	1.95		2.12					

A similar close approximation to a geometric series was found for the distribution of word length in terms of syllable number of Omar Khayam's *Rubaiyat* in Fitzgerald's translation. Both the observed series and the geometric series fitted to it are shown in the following table (after Elderton, "A Few Statistics on the Length of English Words", *J. Roy. Stat. Soc.*, CXII, p. 436).

The possibility of fitting such distributions by means of a geometric series was considered by Elderton to be no more than a fluke of chance. No significance was attributed to it, probably because he felt at a loss to account for such a regularity. An explanation, however, is provided in quite a striking way by the theory of information. We consider the fundamental theorem of that theory:

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 60

Distribution of word-length according to syllable number in Omar Khayam's Rubaiyat (Fitzgerald's translation)

x	1	2	3	4	5	Total
f	1832	420	88	12	5	2357
‰	777	178	38	5	2	1000
g.s.	783	170	37	8	2	

$$-\Sigma p \log p \leqslant M \tag{16}$$

For the sign of equality to apply, the $-\log p$ values must equal the variable, that is symbol lengths, excepting the possibility of exactly compensating deviations. Since the scale of the variable forms an arithmetic series, the $-\log p$ values must form such an identical series, and, consequently, the p's a geometric series, moreover, a geometric series of precisely the type found in our shorthand analysis, viz. with c.f. 0.5 and mean 2.

We shall now state this argument more precisely.

We have by the fundamental theorem

$$H = -\sum_{n=1} p_n \log_2 p_n \leqslant M \tag{17}$$

where p_n is the probability that a "word" contains n syllables (or "strokes"), and M is the mean word length,

$$M = \Sigma n p_n \tag{18}$$

We have equality in (16) if, and only if

$$p_n = 2^{-n}, \; M = 2 \tag{19}$$

If $M \neq 2$, equality is impossible (Herdan, 1953).

The series in Table 59, col. 8 is not very different from the geometric series with common factor .5 and mean 2 (col. 9), nor is the difference of the mean for the combined samples 2.08 incompatible with the assumption of a sampling deviation from the theoretic value of 2.

The conformity of the series of p_n with the particular geometric series of col. 9, Table 59, implies a gradient of p_n such as to ensure ready differentiation of word length, and length of gradient so as to ensure a sufficient range of the variable word length, and thus appears to be *a condition of optimal coding*.

13.2 GENERALISATION OF THE CONDITION FOR OPTIMAL CODING

However, word length may be distributed so as to approximate a geometric series with common ratios other than 1/2 and consequently, with other values than 0.5 for p_1.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

As Good has shown (private communication), the equality (16) is only a special case of maximising H when M is given, namely for $M = 2$. The general formula for the maximal H, M being given, is

$$\max (H \mid M) = M \log M - (M-1) \log (M-1) \tag{20}$$

If M is given and H is maximised, then the probabilities

$$p_n = 1/M\,(1-1/M)^{n-1} \tag{21}$$

form a geometric series with common ratio $(1 - 1/M)$. (19) is a special case of (21). (Appendix *B*.)

It can easily be proved that the right-hand side of (20) cannot exceed M and is equal to M when $M = 2$, and only then. This constitutes a proof of the inequality in (16).

To show this, let $f = M - M \log_2 M + (M-1) \log_2 (M-1)$. Then $f'(M) = \log_2 \{2(1-1/M)\}$. This is < 0 if $M < 2$, $= 0$ when $M = 2$, and > 0 when $M > 2$; so $f > 0$, except when $M = 2$, and then $f = 0$.

This means that geometric series with common factor other than 1/2 could be arrived at by maximising H subject to M (other than 2) being held fixed.

$\max (H \mid M)$ could be used as a measure of efficiency preferable to H/H'.

The geometric series of probabilities is *optimal* because it checks the fall of the gradient, and thus ensures a certain range of the variable. Although efficiency of a code requires H to be less than H', since redundancy is measured by $1 - H/H'$, yet too great a concentration of the range implied in too steep a gradient of probabilities would result in reducing, and even minimising, the number of categories of the variable. The geometric series through maximising H for a given M acts as a barrier to the fall of the gradient, and thus prevents the minimisation of the number of categories in spite of the reduction of H against H'. Arithmetically, this means that an optimal code requires not only a sensible difference between H and H', but also that the relative entropy H/H' should not fall below a certain value; or that $M - H$ should be a minimum as far as this is compatible with the condition that H is sensibly smaller than H'. Consequently, the redundancy for optimal coding represents an upper limit beyond which it would tend to render the code less efficient.

Considering that we have found the geometric series of p_n in poetry (Omar Khayam), in Pitman's shorthand and, as was shown by the author,[1] also in the professional coding of diseases, it would seem that whenever a code for symbolic presentation is not the accumulated result of co-operation by the members of the speech community during centuries, but the product of directed effort by a few or by a single individual with a view to its being convenient, then a geometric series will be aimed at for the p_n. In support of this, we describe below the method for word coding devised by Brillouin.

[1] G. Herdan, "The mathematical relation between the number of diseases and the number of patients in a community", *J.R.S.S.*, Series A, Vol. 120, pp. 320-330, 1957.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

13.3 THE UNCERTAINTY RELATION OF INFORMATION THEORY

The entropy H_{max} for optimal coding stands in an important relation to the variation of both symbol lengths and their probabilities: it prevents the simultaneous decrease of the dispersions of symbol lengths and probabilities beyond the point where the code would become inefficient.

If we minimise the variation (error) of the probabilities p_n by eliminating their gradient (i.e. making the distribution an equi-distant one), $H \rightarrow H'$ and the error in guessing tends towards the maximum for the given situation; if we minimise the variation (error) in length of the coding symbols, x_n, by making the gradient of p_n too steep, $H \rightarrow 0$, and the symbolic expression (language, say) becomes a prolonged moo.

As the variation in p increases, that of x should decrease, and vice versa, if the code is not to become inefficient or even useless.

For optimal coding, the simultaneous minimisation of the gradient of p_n and of the range of x_n must be prevented; in other words, the gradient must be such as to prevent simultaneously $H \rightarrow H'$ and $H \rightarrow 0$. This is achieved by maximising H, for a given M, since this implies a lower limit for H compatible with the value of M as resulting from the distribution of coding symbols. Strictly speaking, H_{max} is a border line value above which $H \rightarrow H'$ too closely, the gradient of p_n becoming too flat, and below which $H \rightarrow 0$ too closely, the gradient becoming too steep for the code to be truly efficient.

This shows paradoxically that optimal coding requires a code structure of the type which is known in physics as the Uncertainty Relation. For this relation it is essential that the product of the variances of the two variables involved should have a lower limit ($h/4\pi$ according to the Heisenberg Uncertainty) which implies that the product of the two variances cannot decrease simultaneously.

The reciprocal relation between the probability of symbols belonging to the coding system and their length is known to be a requirement for efficient coding, but only as an empirical phenomenon. No attempt has been made to establish an intrinsic relation between that reciprocity and information as defined by entropy. Our derivation of the optimal entropy as the value of H which implies a gradient such as to control the simultaneous variation of probabilities and range of symbol length after the manner of the Uncertainty Relation of physics shows the reciprocity relation to be an integral part of the theory of coding, and thus incorporates in information theory the second great statistical principle of theoretical physics. Information theory then shows language to be a form of symbolic expression and communication subject to the two principles having the most general validity in the physical world, the Entropy and the Uncertainty Relation (Vol. I, Sect. 14.3).

13.4 H/H_{max} AS CRITERION OF THE QUALITY OF CODING

(a) *Word length as a case of optimal coding.* Let us consider the results of Sects 10.1,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

10.2, Vol. I in the light of the criteria for optimal coding. We arrived there at the conclusion that in the languages under examination the syllabic device behaved like a binary code (Sect. 10.3); to be more explicit: word length in terms of syllable number satisfied the requirement of efficient coding.

To test this conclusion in the light of the criterion of optimal coding, we shall calculate for each text H_{max}, according to formula (20) and form the ratio H/H_{max} whose deviation from unity indicates the degree of non-compliance with the requirement for optimal coding.

Table 61 lists the results for a selection of texts from Table 62, Vol. I.

TABLE 61

	M	H	H_{max}	H/H_{max}
Bible (1st sample)	1.21	.81	.81	1.00
Bacon	1.41	1.22	1.56	.78
Gray (verse)	1.33	1.05	1.07	.98
Johnson	1.60	1.51	1.53	.98
Modern English (A)	1.20	.76	.78	.97
Goethe	1.74	1.65	1.71	.96
Lichtenberg	1.73	1.65	1.70	.97
Pushkin	2.02	1.83	2.02	.91
Turgeneff	2.24	2.07	2.22	.93
Tolstoy (I)	2.20	2.11	2.19	.96

The smallness of the deviations of H/H_{max} from unity in all texts examined lends support to the conclusion reached before, viz. that word length in terms of syllable number can be regarded as an efficient coding device, as defined in information theory. The slight increase in the deviation from unity as we proceed from English (Bacon is an exception) to German and from there to Russian is also in agreement with the conclusion reached before that "efficiency" in this particular aspect of the language code decreases slightly in that direction.

TABLE 62

		M	H_{max}	H	H/H_{max}
Flaubert	T.68a	2.480	2.41	2.333	.97
	68b	2.305	2.276	2.14	.94
Goethe	T.69a	3.177	2.857	2.81	.984
Lichtenberg	T.69b	3.230	2.883	2.81	.98
Tolstoy	T.71a	4.253	3.349	3.358	1.003
	71c	3.925	3.214	3.253	1.012

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(b) *Word length in translation as a case of bi-variate optimal coding.* From the data as given in any of the Tables 69-73, Vol. I, we can calculate the bi-variate average, i.e. the sum of products: cell frequency × product of the corresponding syllable numbers divided by 1000. Using these bi-variate averages in the formula for H_{max} (20), we calculate the optimal bi-variate entropy (Table 62) and the ratio H/H_{max}.

13.5 BINARY CODING OF WORDS AS A CASE OF OPTIMAL CODING

The concept of optimal redundancy as requiring a geometric series of p_n provides the explanation of the method for word coding devised by Brillouin (1953). Brillouin's object is a system of binary coding of a sample of about 10,000 English words which he considers, following Shannon, a sufficient vocabulary for the English language. The words, as a qualitative variable, are replaced by symbols consisting of combinations of the binary units 0 and 1. This enables one to use a channel capable of transmitting only one type of pulse; no pulse means 0 and one pulse means 1. Since composite signals have different lengths, we must provide some means for recognising the end of a signal. The double, triple, or quadruple 0 may characterise the termination of the pulse system representing a word. This means that we must keep only single 0's in the signal proper. We thus eliminate multiple zeros in the main signal and use them only in the termination.[2] Every signal starts with a 1 and ends with 00, 000, 0000 .... The first such combinations are shown in Table 63.

TABLE 63

I	II	III	IV	V
1.00	10.00	100.00	1000.00	10000.00
	11.00	101.00	1010.00	10100.00
		110.00	1011.00	10101.00
		111.00	1100.00	10110.00
			1101.00	10111.00
			1110.00	11000.00
			1111.00	11010.00
				11011.00
				11100.00
				11101.00
				11110.00
				11111.00

In the fourth column, the combination 10001 has been omitted because of the multiple zeros. For the same reason we have omitted 10001.00, 10010.00, 10011.00, and 11001.00 from the fifth column.

In the following discussion we shall omit the last two zeros, which appear at the

[2] I.e., no signal is to consist of zeros only.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

end of every signal. Let us suppose that we have found all the signals of a certain column (for instance, the seven signals of column IV, with four digits in the main part of each signal). Let

n be the rank of the column,
$N_{n,1}$ be the number of signals ending with 1,
$N_{n,0}$ be the number of signals ending with a single 0,
and $N_{n,00}$ be the number of signals ending with a series of zeros.

For instance, in the case of volumn IV (ignoring the last two zeros after the point),

$$n = 4,\ N_{4,1} = 3,\ N_{4,0} = 2,\ N_{4,00} = 2$$

We now want to compute the numbers $N_{n+1,1}$, $N_{n+1,0}$, and $N_{n+1,00}$ for signals with $n+1$ digits in the main part. The $N_{n,0}$ and $N_{n,1}$ signals can be completed with either 0 or 1. The $N_{n,00}$ can be completed only with 0, and hence we have the following relations:

$$\begin{aligned} N_{n+1,1} &= N_{n,1} + N_{n,0} \\ N_{n+1,0} &= N_{n,1} \\ N_{n+1,00} &= N_{n,0} + N_{n,00} \end{aligned} \tag{22}$$

These relations yield, step by step, the results shown in Table 64.

We can use the relations (22) to compute the relative probabilities p_{n1},, p_{n0} p_{n00} of signals whose main part ends with 1, 0, or 00.

$$N_{n1} = N_n \cdot p_{n1},\ N_{n0} = N_n \cdot p_{n0},\ N_{n00} = N_n \cdot p_{n00},\ N_n = N_{n1} + N_{n0} + N_{n00} \tag{23}$$

Equations (22) and (23) together determine the successive values of the p's. When n is large ($n > 10$, for instance), these probabilities p become independent of n and one easily obtains

$$p_1 \sim p_{00} = \tfrac{1}{2}(3 - \sqrt{5}) = 0.382,\ p_0 = 0.236 \tag{24}$$

and eqs. (22) result in

$$N_{n+1} = N_n(2p_1 + 2p_0 + p_{00}) = 1.618\, N_n; \tag{25}$$

hence

$$N_n = D_o(1.618)^n \text{ with } D_o = 1.163$$

$$N_1 + N_2 + \ldots + N_n = D_1(1.618)^n \text{ with } D_1 \sim 3 \tag{26}$$

These approximate relations can be checked with the exact figures of table 64 and are found to represent a good approximation for $n > 10$.

13.6 OPTIMAL CODING AND FIBONACCI NUMBERS

The resulting series of frequencies as a geometric series with common factor 1.618 is also known under the name of the Fibonacci Series. This means that this type of coding, which Brillouin has devised empirically as the most economical requires, as

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 64

Binary Coding of Words

	Column number, n									
	1	2	3	4	5	6	7	8	9	10
N_{n1}	1	1	2	3	5	8	13	21	34	55
N_{n0}	0	1	1	2	3	5	8	13	21	34
N_{n00}	0	0	1	2	4	7	12	20	33	54
ΣN	1	2	4	7	12	20	33	54	88	143
Total number of words coded	1	3	7	14	26	46	79	133	221	364

	Column number, n									
	11	12	13	14	15	16	17	18	19	
N_{n1}	89	144	233	377	610	987	1597	2584	4181	
N_{n0}	55	89	144	233	377	610	987	1597	2584	
N_{n0}	88	143	232	376	609	986	1596	2583	4180	
ΣN	232	376	609	986	1596	2583	4180	6764	10945	
Total number of words coded	596	972	1581	2567	4163	6746	10926	17690	28635	

we predicted on theoretical grounds, a geometric series. The definition of optimal redundancy given above thus explains why Brillouin should have constructed his series as he did. To show that it fulfils the condition for optimal coding we imagine the words to be distributed in groups of decreasing size—the ratio of successive group sizes being 0.618—and let the coding symbols start with unity (that is we count only the number of units in front of the period).

This gives the series

	for the group series ΣN	and *for the word totals*
1	10945	28655
2	6764	17696
3	4180	10936
.	.	.
.	.	.
.	.	.
19	1	1

which are sensibly identical with the corresponding series in table 64.

We then obtain for the ΣN series

$$M = 2.611 \qquad H = 2.525$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\max(H \mid M) = \log_2 M^M - \log_2 (M-1)^{M-1} = \log_2 \left(\frac{M^M}{(M-1)^{M-1}}\right) = 2.52$$

This shows the Fibonacci Series which Brillouin devised for word coding in an empirical way as the most economical that can be obtained practically, to fulfill the condition set up on theoretical grounds, viz. that H should be a maximum for a given M.

We can now improve upon the general condition for optimal coding. If it is desirable to obtain distinctions in terms of binary units between not only word groups but individual words, then the geometric series which is a condition for optimal coding in general should be the series of the Fibonacci numbers, or more generally, a geometric series with common factor 0.618, in reverse.[3]

[3] For the properties of the Fibonacci series and references, see Vol. I, p. 129, and also Whittaker and Watson (1935)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIV A

AN INEQUALITY RELATION BETWEEN YULE'S CHARACTERISTIC K AND SHANNON'S ENTROPY H

14.1 ENTROPY AND LINGUISTICS

If the entropy is used in linguistics in the form and with the interpretation given to it by communication engineers, as so and so many "bits of information", it requires much additional explanation, and even then confusion is not always avoided. For one thing, it may be now regarded as established that the entropy H measures not the information, but rather the lack of it, and that, therefore, the negative entropy—the negentropy—should be used for measuring information; secondly, H is a logarithmic quantity, and as such not readily interpretable in non-logarithmic terms; more especially the "bits of information" have not really a linguistic meaning.

The transformation which we shall propose takes these shortcomings of the traditional expression of entropy into account, and by doing so changes it into an easily interpretable statistical quantity of general applicability and of interest to linguists—the *repeat rate*. It is based upon an inequality relation between the two parameters which are indigenous to literary statistics: Shannon's entropy $H = -\Sigma p_i \log_2 p_i$, where p_i is the probability of occurrence of the symbol i, and Yule's characteristic (apart from a factor 10^4 to avoid decimals)

$$K = \frac{S_2 - N}{N^2}$$

where
$$N = \Sigma r\, n_r \quad \text{and} \quad S_2 = \Sigma r^2 n_r \tag{32}$$

and where n_r is the number of different words occurring with the frequency r.

14.2 A TRANSFORMATION OF THE ENTROPY—THE INEQUALITY RELATION BETWEEN K AND H

The negative entropy $-H = \Sigma p_i \log_2 p_i$ is the arithmetic mean of the logarithmic variable $\log_2 p$. Its antilog is, therefore, the *geometric mean* G of p_i. In symbols:

$$\text{Antilog}\,(-H) = G \tag{33}$$

The *arithmetic mean* A of p, as a weighted average, or the expectation of p, is

$$A = \Sigma p_i p_i = \Sigma p_i^2 \tag{34}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Σp_t^2 is the repeat rate, i.e., the probability that two particular linguistic events, say words, selected at random from the text under consideration should turn out to be the same word.

Using the inequality between arithmetic and geometric mean

$$G \leqq A \tag{35}$$

we obtain the following inequality between the entropy and the repeat rate of particular events:

$$\text{Antilog}\,(-H) = G \leqq A = \Sigma p_t^2 \tag{36}$$

On the other hand, it can be shown that Yule's characteristic K is an unbiased estimate of the sum of the repeat rates for particular linguistic events Σp_t^2. (For the proof, see Appendix D).

We write therefore: $$\text{Antilog}\,(-H) \leqq K \tag{37}$$

In words:

The repeat rate or the Characteristic is at least equal to the antilog of the negentropy, or the antilog of the negentropy is a lower limit of the repeat rate or of Yule's K.

The following illustration is based upon the data on the relative frequency of phonemes in literary English compiled by M. Fowler (Section 5.5).

From the values for v_m we calculate the corresponding value of K^* according to the formula

$$K^* = v_m^2 + 1/N \tag{38}$$

We thus get from Table 24, last col. of partial tables:

$$K^*_I = .0490$$
$$K^*_{II} = .0509$$
$$K^*_{III} = .0522$$

These values are sufficiently close for us to regard the Characteristic K as stable for the frequency distribution of English phonemes, and thus as a constant regardless of style (see remark in Introduction to Part II about Trubetzkoy's view on this matter). For applying the established Inequality Relation we choose K^*_{III}.

The entropy $H = -3.322\, E(\log p)$ results for the frequency distribution of phonemes in Table III of Fowler's paper as

$$H = 4.5041$$

from which we have

$$\text{Antilog}\,(-H) = .0441$$

which fulfills the condition of inequality, formula (37):

$$\text{Antilog}\,(H-) = .0441 \leqslant K^* = .0522.$$

Using the interpretation of K as a repeat rate, we can express this by saying that a pair of phonemes picked at random from a modern English text will turn out to be the same phoneme about 5 times in 100, or 1 in 20, and at least 4.4 in 100, or 1 in 23.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The relation between the interpretation of K as the repeat rate and as the coefficient of variation of the mean is as follows:

In formula (32) both r and n_r are frequencies, of word occurrences and vocabulary respectively, and may by division through the corresponding totals be transformed into relative frequencies. For great samples, these relative frequencies are sensibly stable and may be regarded as estimates of the corresponding probabilities,

$$p_r = r/N \text{ and } \pi_r = n_r/n \quad (n = n_1 + n_2 + \ldots n_s).$$

Using the first transformation, we have

$$K^* = \frac{S_2}{N_2} = \sum_{r=1}^{s} \frac{r^2}{N^2} n_r = \sum_{r=1}^{s} p_r^2 n_r$$

and summing over the individual p_r instead of over their groups n_r, we may write

$$K^* = \sum_{\mu=1}^{n} p_\mu^2$$

which shows K to be the repeat rate of particular words.

For using the second transformation, we write

$$K^* - \frac{1}{n} = \frac{S_2}{N_2} - \frac{1}{n} = \frac{n \Sigma r^2 n_r - [\Sigma r n_r]^2}{n [\Sigma r n_r]^2}$$

Dividing numerator and denominator by n^3, the expression on the right becomes

$$\frac{\{\Sigma r^2 \pi_r - [\Sigma r \pi_r]^2\}/n}{[\Sigma r \pi_r]^2} = \frac{\sigma_r^2/n}{M_r^2}$$

which shows K, for great n, to be equal to the coefficient of variation of the mean of the variable r.

Relation (37) is of great importance both practically and theoretically. From the practical angle, it enables us to obtain an estimate in terms of a lower limit of the repeat rate if K cannot be calculated because the variable is a qualitative one. Furthermore since, as a rule, H is easier calculated than K, it provides a method for obtaining quickly a lower limit for the latter. From the theoretical angle, it provides the sought-for correct linguistic interpretation of the entropy in terms of the repeat rate of linguistic forms.

In Section 13.2 we derived the value of H for optimal coding as

$$H_{\max} = \log_2 \left(\frac{M^M}{(M-1)^{M-1}} \right) \tag{38}$$

and this value may be used for a critical appreciation of the repeat rate of linguistic forms in a given text. We know by formula (37) that the antilogarithm of $-H$ is a lower limit of K for a given literary text. This does not necessarily imply that an

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

observed K must stand in the same relation to $-H_{max}$. But should it be so it would indicate that the repeat rate system of the text in question was consistent with optimal coding.

It follows that the ratio

$$\text{Antilog}\,(-H_{max}) \,/\, K$$

can serve as a criterion of the *optimal repeat rate*: if the inequality holds, the repeat rate is such as to be consistent with optimal coding.

14.3 APPLICATION (UNI- AND BI-VARIATE)

Using the inequality, we can express, for instance, the difference in the entropy for length of English words in terms of number of letters $H = 2.628$ and that for word length in terms of sounds (phonemes) $H = 2.274$ by saying that the repeat rate for specified word lengths in English is at least 1 in 6.3 words, if word length is measured in terms of letter number, and 1 in 4.8 words, if word length is measured in phoneme number. In other words, picking two words at random from a very long text and examining them for word length, we shall find that word length repeats itself slightly more often if measured in terms of phonemes.

Language translation may be regarded as bi-variate coding of messages.[1] If the entropy is calculated for a given quantitative characteristic, such as word length, for each of the two languages, original and translation, independently, the values so obtained will differ, as a rule, among themselves and also from the value of the bi-variate entropy, that is the entropy calculated from the word length of translational equivalents (words denoting the same concept in the two languages). More precisely, the conditional entropy, i.e. the entropy for word length in translational equivalents of words in the original of a specified length is less than the unconditional entropy. Thus for translation from French into English we find the former as 0.923 bits and the latter as 1.410; in translation from German into English the former is 1.162 bits and the latter 1.648; in translation from Russian into English the former is 1.153 bits and the latter 2.100 (Sect. 11.4 Vol. I).

These relations are not too clear as long as they are expressed by means of the entropy in its conventional form. The inequality relation enables us to make the difference between the conditional and the unconditional entropy more readily understandable. In terms of that relation we express the differences as follows: for French-English translation the repeat rate of word length in the French original is at least 1 in 2.6, or 1 in 3 approximately, and rises in the translational equivalents to at least 1 in 2; for German-English translation the repeat rate of word length in the original is also 1 in 3 approximately, and in the translational equivalents rises again to at least 1 in 2 approximately; for Russian-English translation the repeat rate of word length in the original is 1 in 4 approximately, and rises in the translational equivalents to 1 in 2 (2.4 precisely). For another application see p. 303.

[1] Vol. I, Chpt. 11.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIV B

MARKOFF-CHAIN MECHANISM IN LANGUAGE TRANSLATION

14.4 THEORY

It was said by E. Schrödinger of wave-equation fame (1946)[1] that it is "one of the fascinating features of statistical thermodynamics (which provided the prototype of statistical information theory) that quantities and functions, introduced primarily as mathematical devices almost invariably acquire a fundamental physical meaning".

It is my contention that quantities and functions of information theory, introduced primarily as mathematical models for the transmission of coded messages, acquire often a fundamental linguistic meaning, e.g. the interpretation of the Entropy as a measure of efficiency of the linguistic code (Vol. I, Chpt. 10). We shall now give the linguistic interpretation of the information-theoretical concept of "noise". This is the name for any interference with the sound pattern of the original message during transmission. Now, noise as such has no place in language structure, and the question arises whether that part of the theory in which "noise" plays an important role is applicable to language structure at all. In other words: can we give that concept a linguistical interpretation? Unfortunately, language structure is not a field to which many experts in information theory would turn their attention, which is understandable. What is less understandable is that they should turn a completely blind eye to the interpretations which become necessary when the theory is applied to language structure. This is one of the reasons for the unending misunderstandings between the information mathematician and the mathematical linguist.

In accordance with Schrödinger's remark, it is the mathematical device for the interference by "noise" which points the way. That device consists in regarding the original message and the form in which it is received under "noise" interference as being correlated, more precisely, as standing in the relation of the members of a Markoff chain (Sect. 10.3). The bi-variate character of that relation recalls that between original and translation of a text. A translation can be regarded as an instance of double coding: the identical content is expressed in two different language codes. It is first encoded in the "original" language from which the translator has to decode it previous to encoding it in that of "translation". For our purpose, there are two types of characteristics to be considered. With regard to some language features,

[1] E. Schrödinger, *Statistical Thermodynamics* (Cambridge, 1946).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

such as the particular sequence of phonemes within the words, the target language may show little or no similarity with the source language according to the degree of genealogical relationship. With regard to other features, however, there may exist a more or less obvious dependence of the forms of translational equivalents. Such a feature is, for instance, word length in terms of syllable or phoneme number (Vol. I, Sect. 11.3). If the connection between the word-length of translational equivalents can be satisfactorily represented by the mathematical model for the transmission of messages under noise, the conclusion would be justified that what corresponds to noise in language structure is the stochastic change of certain quantitative characteristics of linguistic forms, such as word-length, in translation.

The definition of the Markoff chain between two events, X and Y, in a linear sequence of repeated occurrences of these events is that in addition to the probabilities for each event occurring independently, there are transition probabilities if the occurrence of the event X is dependent upon whether the immediately preceding event is X or Y, and similarly for the event Y; the part of the total sequence, however, preceding a particular event, X or Y, has no influence upon the occurrence of the event.

We consider a linguistic text. For the purpose of this investigation we regard the words as units of which the text has been built up. The translation of the message into another language is a similar text whose units are equivalents of those in the original.

Let the units or elements of the original text be

$$a_1, a_2, a_3 \ldots$$

having respectively the properties

$$x_1, x_2, x_3 \ldots$$

the subscript denoting the order of occurrence in the text, and let the equivalent units of the translation be

$$\alpha_1, \alpha_2, \alpha_3 \ldots$$

having respectively the properties

$$y_1, y_2, y_3 \ldots$$

the subscript denoting here not the order in the translation, but the correspondence with the units of the original.

If the text represented a truly random sequence of the values x and y, then the combination of the elements in pairs, triplets, etc. should have no effect upon the randomness of their arrangement. If, on the other hand, the characteristic y of the translation was to some extent dependent upon the characteristic x of the original, a Markoff chain would result as the following consideration shows. We conceive of the original text in the translation as being brought into a one-to-one correspondence, a_t corresponding to α_t which results in the following series of doublets:

$$\alpha_1, \alpha_1/\alpha_2, \alpha_2/\alpha_3, \alpha_3 \ldots\ldots$$
$$x_1, y_1/x_2, y_2/x_3, y_3 \ldots\ldots$$

That is, we have conceived of the original and the translation as being combined into

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

one linear sequence of units, in which translational equivalents follow immediately upon one another, with transition probabilities between particular values of x and y as given in the p_{ij} matrix, where the subscripts of p refer no longer to the order in the text but to the value of the variables x, y:

	x_1	x_2	x_3 ....
y_1	p_{11}	p_{12}	p_{13} ....
y_2	p_{21}	p_{22}	p_{23} ....
y_3	p_{31}	p_{32}	p_{33} ....

If such transition probabilities exist, the occurrence of α_i having a particular y_i value will, to that extent, be dependent *only* upon the preceding a_i having a particular x_i value, but not upon the whole preceding array of units, and the sequence would thus fulfil the conditions for a Markoff process.

As the criterion of dependence between a and α with respect to the properties x and y we use the relation between the bi-variate entropy, as defined in Information Theory, $H(x,y)$ and the uni-variate values $H(x)$, $H(y)$. Let $p(i,j)$ be the probability of the joint occurrence of the events according to the above matrix. The entropy of the joint event is then calculated as

$$H(x,y) = -\Sigma p(i,j) \log_2 p(i,j) \tag{40}$$

while those of the single event are given by

$$\begin{aligned} H(x) &= -\Sigma p(i) \log_2 (i) \\ H(y) &= -\Sigma p(j) \log_2 (j) \end{aligned} \tag{41}$$

In the case of the stochastic independence of x and y, the bi-variate entropy $H(x,y)$ equals the sum of the uni-variate entropies $H(x)$ and $H(y)$:

$$H(x,y) = H(x) + H(y) \tag{42}$$

This property of H suggests using the difference

$$R = H(x) + H(y) - H(x,y) \tag{43}$$

as a criterion for the dependence or otherwise of the translation upon the original as regards the chosen characteristic. It measures the constraint which the original exerts upon the translation in that respect. The difference R may also be written in another way which shows it to be the result of both, probabilities and transition probabilities:

$$R = H(y) - H_x(y) = H(x) - H_y(x) \tag{44}$$

where $H_x(y)$ is the transition probability of values of y for any given value of x, and $H_y(x)$ the transition probability of values of x for any given value of y. Since it is based upon the transition probabilities of x and y in immediate neighbourhood only, its mathematical structure is indicative of a Markoff chain between the two groups of events, i.e. word lengths in the original and in the translation.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

14.5 APPLICATION

As a linguistically important and comparatively easily measurable property with regard to which to compare the original with the translation, the number of syllables per word was selected. As experimental material for the investigation were chosen translations into English from four languages; French, German, Czech and Russian. The bi-variate distributions required for the computation of the various entropies were obtained by recording for each word occurrence its length in terms of the number of syllables in the original and in the language of translation.

Table 65 is an example of the bi-variate distribution and calculation of the relevant quantities required for establishing a Markoff chain. The results from 8 such bi-variate syllable counts are listed in table 66 (Vol. I, Tables 68-72).

TABLE 65

Translation from Tolstoy (War & Peace) (N = 1,106, reduced to N = 1,000)

		No. of syllables in Russian words							Totals
		1	2	3	4	5	6	7	
No. of syllables in English words	1	305	197	84	26	2	1	—	615
	2	28	108	79	38	16	2	—	271
	3	2	17	30	21	9	2	1	82
	4	1	—	4	9	7	3	1	25
	5	—	—	1	2	2	1	1	7
Totals		336	322	198	96	36	9	3	1,000

Russian
$M_R = 2.21$
$H(R) = 2.10$

English
$M_E = 1.54$
$H(E) = 1.42$

$H(R,E) = 3.253$

and thus $H(R,E)\ H(R) + H(E) = 3.52$

$R = H(R) + H(E) - H(R,E) = .27$

From the fact that the bi-variate entropy is in each case less than the sum of the uni-variate entropies, we draw the conclusion that *in translation from one language into another there exists a stochastic dependence of the Markoff process type* between certain quantitative characteristics of the words in the original and in the translation.

The strength of the Markoff chain is measured by the difference (R) between the bi-variate entropy for the characteristic of "word length" and the sum of the uni-variate entropies for the same characteristic.

Different pairs of languages are characterised by different values of R, denoting

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 66

French-English	H(E)	H(F)	H(FE)
Flaubert, Sample I	1.21	1.41	2.14
Sample II	1.17	1.31	2.33
German-English	H(E)	H(G)	H(GE)
Goethe	1.46	1.65	2.81
Lichtenberg	1.48	1.65	2.81
Czech-English	H(E)	H(C)	H(CE)
Masaryk, Sample I	1.63	1.88	3.14
Sample II	1.57	1.96	3.11
Russian-English	H(E)	H(R)	H(RE)
Tolstoy I	1.42	2.11	3.36
Tolstoy II	1.55	2.21	3.25

different intensities of the Markoff chain between the original and translation as regards word length.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XV

THE SKEW DISTRIBUTIONS OF STATISTICAL LINGUISTICS

15.1 RECIPROCITY RELATIONS OF LINGUISTIC FORMS

The development of quantitative linguistics is characterised by the fact—perhaps not unusual in the development of science—that certain far-reaching regularities or similarities of word- (phoneme-, morpheme- etc.) counts have passed more or less unnoticed, or, if noticed, failed to arouse even curiosity, let alone surprise. And yet, it is these similarities which may be said to pave the way for establishing linguistics as a quantitative science.

There are two such similarities. One is the similarity between the series of relative frequencies for one *particular* type of linguistic form, such as phonemes, letters, morphemes, words etc., or briefly the stability of relative frequencies of linguistic units. This has led to the interpretation of de Saussure's "langue-parole" dichotomy as that between the statistical population and the sample.[1] The other is the similarity in form between the frequency distributions of *different* types of linguistic units. The distribution of phoneme occurrences according to the phonemes arranged in descending order of frequency, that of word occurrences according to the words arranged in descending order of their rank, of vocabulary items according to the frequency of occurrence, of words according to their length in terms of the number of syllables (phonemes, letters) and of sentences according to the number of words, all show the characteristic reversed J-shape.

As said above, this has not passed altogether unnoticed; linguists have tried to account for the similarity of some of these curves by the—obvious—relationship of the variables concerned, without however attaching much importance to it. Mathematicians, too, have noticed the similarity and have also drawn attention to the similarity between linguistic distributions in this respect and certain distributions of Biology (number of genera according to the number of species, relative abundance of species in nature), of Economics (income distributions) and Demography (city sizes). This heterogeneity of material led one investigator (Simon, 1952) to conclude that the apparent uniformity must be due, not to a particular linguistic property, but only to a common probability mechanism. Such a mechanism he derived under relatively weak assumptions, with, consequently, a certain generality of application. However, as will

[1] Vol. I, Sect. 5.3.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

be shown, the mathematical model he constructed for that mechanism cannot be regarded as a satisfactory generalisation of the different reciprocity relations. It will become apparent that the reciprocity relation which manifests itself in all these distributions is an essential functional characteristic of symbolic expression in general, and of linguistic coding in particular.

For the purpose of the discussion it is advisable to distinguish between linguistic distributions in which the reciprocity relation within the variables is brought about artificially, i.e. by making one variable the inverse function of the other, and distributions in which both variables are genuine linguistic quantities.

15.2 ARTIFICIAL RECIPROCITY RELATIONS

15.2.1 The distribution of frequency according to symbol length

The units of the information-theoretical measure are binary units ("bits"). We say that information is measured by the entropy H as so and so many "bits". This means that if n objects are arranged in a certain order and the series then divided into sub-series of $n/2$ objects, $n/4$, $n/8$ etc., a particular object can be identified by H (rounded off to the nearest unit) questions as to whether it lies within the upper or lower part of the whole series, the upper or lower part of half the series in which it has been located by the previous question, and so on.

The "bits" of information are no doubt an artificial variable since they imply a particular arrangement of the data, such as does not occur in nature.

In order now to make the code more efficient or economical, there should be a certain co-ordination between symbol length in terms of "bits" and frequency of occurrence of symbols. The basic principle in this respect is that the symbol occurring most frequently should be the smallest in terms of "bits", the second most frequent occurring symbol longer, and generally that the length of symbols should be inversely related to the frequency of occurrence. Not all codes satisfy that principle. The Morse Code, for instance, although working with a two-element code, dot and dash, has not applied the economical principle systematically (see Vol. I, Sect. 7.1).

In an efficient code we expect, therefore, a reciprocity relation of some kind between symbol length and occurrence probability of symbols.

As to the particular mathematical law governing the frequency distribution of symbols according to their length in terms of "bits", we know from Chapter XIII that for an optimal code the frequency gradient is such as to satisfy the requirements of a geometric series. This means that in such a code there exists a reciprocity relation between symbol length and occurrence frequency such that on semi-logarithmic paper the function will plot as a straight line.

Although the information-theoretical variable is essentially an artificial one, designed with a view to the transmittance of information by electrical applicances, yet it is easily understandable that in an efficient natural code, such as language might be

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

expected to be in some respects, the relation between symbol length and occurrence frequency might be very close to the information-theoretical idea. As we know from Chapter XIII, this is actually the case for symbol (word) length in terms of number of letters, phonemes and especially number of syllables.

15.2.2 The distribution of frequency according to rank

1

It was pointed out (Section 1.6) that the Zipf law is not a law in the correct and strict sense of the term. To the extent, however, to which it is true, that is as far as the bi-logarithmic straight line law fits the frequency of words, it may be used for deriving a characteristic of such distributions, which is important for the mechanical handling of word masses, especially with a view to mechanical translation. The Zipf law may, therefore, be said to have its uses in the technology of language.

If the most frequent word in a given sample is numbered 1, the second most frequent 2, and so on, the *rth* most frequent word is said to have rank r. If the frequency of the *rth* word is f_r, then according to the Zipf law

$$rf_r = C \tag{45}$$

where C is a constant, from which it follows that the occurrence number of the *rth* word is

$$f_r = C/r$$

The total number of word occurrences N of n different words is then calculated as the integral from $r = 1$ to $r = n$

$$N = \int_{r=1}^{n} \frac{C\,dr}{r} = C \log_e n \tag{46}$$

Equation (46) represents the area under the curve when f_r is plotted against r.

We shall also require the first moment of the distribution

$$\int_{1}^{n} \frac{r\,C\,dr}{r} = C(n-1) \sim Cn \tag{47}$$

for great n.

Since N represents the total number of occurrences, it follows that division of N by C should give $\log_e n$ and, alternatively, division of N by $\log_e$ should give C. On the other hand, C can be obtained empirically from the frequency distribution, and the comparison of the two values of C would provide evidence for and against the correctness of the assumption of the bi-logarithmic law.

That law implies, strictly speaking, a one-to-one correspondence between rank and frequency, i.e. that there is exactly one word having rank r. In actual texts there are, as a rule, several words with the same frequency—which is indeed the outstanding

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

characteristic of the word count distribution—and which must be given the same rank number. In this case the Zipf law does not strictly apply, and it is therefore uncertain to what extent the theoretical relation between H, n and C will agree with observation. The following paragraph will give an answer to this.

2

We shall now test the theoretical conclusion about the relation between N, n and C, using some of the extensive material given by R. Morgenthaler in his statistical dictionary of New Testament Greek (1958).

TABLE 67

N.T. Book	Occurrences (N)	Vocabulary (n)	$C = N/\log_e n$	$C' = rf_r$ (from observations)
Matthew	18278	1691	2458.2	2650
Mark	11229	1345	1558.7	1600
Luke	19404	2055	2543.2	2400
John	15420	1011	2228.2	2650
Acts	18374	2038	2411.2	2650
Paul	32303	2648	4098.4	4250
Hebrew	4942	1038	711.6	700
Revelation	9818	917	1439.3	1450
Total	137328	5436	15966.3	15650

In accordance with the empirical fact, that it is the medium range of the distribution for which the Zipf law holds, C' was calculated from the observed values as the product of the rank r which corresponds to $f_r = 50$, for all books.

The agreement between C calculated from the observations in this way (col. 5, table 67) and under the assumption of the bi-logarithmic law (col. 4) is reasonably good. Considering that the distributions are such as to have an increasing number of vocabulary items per frequency class with increasing r, the good agreement between theory and observation shows the applicability of our deduction to actual word counts. This interpretation of the quantity $1/\log_e n$ shows it to be of considerable interest from the machine-technological viewpoint, (para. 3) as well as from the theoretical (para. 4).

3

This result, as observed above, has an important bearing on the mechanical handling of masses of words.

Let us first assume that the dictionary of words is stored in the machine alphabetically and in ascending order of numerical magnitude, i.e. in terms of the number of letters per word (in consecutive positions of the machine store).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The unknown word is then sent to the accumulator of the computor and compared with each of the dictionary entries from the numerically greatest backwards. It can be shown that this method of dictionary search involves, on the average, the examination of $n/2$ entries to locate or identify the unknown word.

From the stand-point of frequency of occurrence of individual vocabulary items, the arrangement according to word length is, to all intents and purposes, a random arrangement. Let us now assume that the arrangement of the entries is systematic according to frequency of occurrence in descending order of frequency so that the most frequent word has rank 1, the second most frequent word rank 2, and so on. If in such a dictionary, that is one in which words are arranged in order of decreasing frequency and increasing order of rank, the look-up procedure is one of successive comparison, the word of rank r will require r look-up operations, and since this word occurs—the Zipf law assumed—C/r times, the total number of look-up operations required to locate a word is C. Thus, for n words contained in the dictionary, nC look-up operations will be required.

On the other hand, we know that for the Zipf law the total number of occurrences (the text length in terms of word number) and thus the total number of words to be searched, is given by

$$N = \int \frac{C\,dr}{r} = C \log_e n \tag{48}$$

It follows that the average number of look-up operations per word is

$$A_n = nC/C \log_e n = n/\log_e n \tag{49}$$

Considering that the total of rank units or the first moment of the distribution according to rank is nC, the ratio A_n is seen to represent the average rank of a word in the given frequency distribution.

This compares favourably with the $n/2$ look-up operations which would be needed under the scheme described above, which makes no use of the frequency element.

4

Formula (49) written in the form

$$A_n \sim n/\log_e n \tag{50}$$

is well known to mathematicians since it represents one of the most remarkable discoveries in the whole of mathematics: the *Prime Number Theorem*. This theorem states that the number of primes to be expected among n integers equals $n/\log_e n$. From an empirical study of prime number tables, Gauss observed that the ratio A_n/n is approximately equal to $1/\log_e n$ and that this approximation appears to improve as n increases. On the basis of such empirical evidence Gauss makes the conjecture that the ratio A_n/n is asymptotically equal to $1/\log_e n$

$$A_n/n \sim 1/\log_e n \tag{51}$$

That $\sim$ cannot be replaced by the ordinary sign of equality is clear from the fact that

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

while A_n is always an integer, $n/\log_e n$ is not. The formal identity of (49) and (50) is, of course, due to the fact that the empirical relation between the variables are in both cases supposed to be well represented by an equi-lateral hyperbola, which on a double-logarithmic grid plots as a straight line.

We may, therefore, express the relation $N/C = \log_e n$ (para. 3) as follows. For a frequency distribution of n vocabulary items, the *total number of occurrences N (text length) divided by the average product $C = rf_r$ equals the reciprocal probability of a prime among n integers*; and the relation $A_n = n/\log_e n$ we may express as follows. For a frequency distribution of n vocabulary items, *the first moment of the rank distribution nC divided by the text length C $\log_e n$ equals the number of primes among n integers.*

In terms of the mechanical handling of word masses with n different vocabulary items, this means that the arrangement of the linguistic material according to the frequency of probability of occurrences has the effect of reducing the number of look-up operations per word to the number of primes among n integers, as against $n/2$ trials per word, if the material is arranged alphabetically or with regard to word length. We may also express this by saying that using the frequency element in the search for words as described has the effect of making word recognise word, if there are n different words, after $n/\log_e n$ trials, or after as many trials as there are primes among n integers. The quantity $n/\log_e n$ is comparable to the entropy $-\Sigma p \log p = \Sigma p \log 1/p$ which represents the number of binary decisions required for identifying a word (or any linguistic form). This supposes the continued halving of the whole sequence of words until the word is located.

In both cases it is the arrangement of entries and with it the method of searching which has the effect of reducing the number of look-up trials against what that number would be without such systematic arrangement.

15.3 NATURAL RECIPROCITY RELATIONS

Simon aims at explaining various linguistic (and other) skew distributions by a uniform stochastic process. He regards their form as the stationary solution of such a process, and believes that these distributions are better fitted by his formula than by the negative binomial or Fisher's logarithmic series, both of which had been suggested for describing their mathematical structure.

The empirical distributions in question are according to Simon: word counts that have been made for English and other languages, e.g. by Yule (1943), Zipf (1949); the so-called Yule-distribution of biological genera according to the number of species they contain; the distribution of writers as contributors to scientific journals according to the number of papers contributed by them; the distribution of cities in a country according to the number of inhabitants; the Pareto distribution of persons according to income.

It is our contention

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(a) that Simon's claim to have provided a uniform mathematical explanation of these distributions rests upon an insufficient realisation of the differences in form between the distributions, and suffers from a neglect of considering the relations between some of them which makes it highly unlikely, if not mathematically impossible, that one mathematical model should fit them all, and

(b) that the assumptions under which Simon's formula is derived do not fit all these distributions equally well, and are sometimes in open contradiction with the material.

15.3.1 Relations between some important skew distributions of linguistics

Taking first the *Zipf distribution* and the *Yule word count*, the former is the distribution of frequencies according to rank, i.e. the frequencies ordered according to magnitude starting with the word of highest frequency as rank 1, the word with the second highest as rank 2, and so on; the latter is the distribution of vocabulary items (words) in a text sample according to their frequency of occurrence. In the Zipf distribution there corresponds exactly one word to a given frequency, and thus to a given rank, and Zipf claims a *bi-logarithmic straight line* relation between frequency and rank. For the Yule word count, on the contrary, it is characteristic that the numbers of words having the same frequency of occurrence form a descending series, and Yule has shown it to be satisfactorily fitted by the *compound Poisson law* which can also be described as the *negative binomial*, whose applicability Simon denies; these distributions may also, and perhaps still better, be fitted by the *lognormal law*, as was shown in Chapter III.

Considering next the *distribution of genera according to the number of species* (which Simon calls the Yule distribution, though to be precise, it was C. Willis who drew Yule's attention to the *bi-logarithmic* property of such distributions), it is interesting to note, and very much to the point I wish to make, that Simon expresses surprise at Yule having apparently forgotten all about it when writing his book about word counts (1943). The explanation is quite simple: the two distributions have little in common mathematically, which makes it unlikely that they should be governed by the same law. The word count distribution is one of linguistic forms (words) according to their frequency of occurrence; that of genera according to the number of species is one of wider concepts according to the number of narrower concepts they contain. Moreover, Yule did not conceive of the latter distribution as one of linguistic forms at all, but as the result of the biological generating process with time of both species and genera within a biological family of animals.

Analogously with the double role played by vocabulary items as the *variable* of the Zipf law and as the *frequency* of the Yule word count, which are the inverse or dual of one another (Vol. I, Sect. 12.3.1), we have in this field the double role played by the number of species as the *variable* of the distribution of genera according to the number of species and as the *frequency* of the distribution of species according to the number of individuals they contain. The former is fitted, approximately, by the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

bi-logarithmic straight line law and more satisfactorily by the Yule distribution law; the latter is fitted satisfactorily by the *logarithmic series*. This has been shown to be also true for the distribution of diseases (using the diagnostic terms of the International Code of Diseases) among the members of a community (Herdan 1957).

We can agree with Simon when he claims that because Yule's paper of 1924 predated the modern theory of stochastic processes his derivation of the law governing the genera-species relation was necessarily more involved than the one which Simon employs, but we must disagree with the claim that his derivation applies uniformly to all the distributions under consideration.

15.3.2 The inadequacy of Simon's assumptions

Simon's stochastic model is based on certain assumptions:

I. The probability that the $(k + 1)$th word is a word that has already appeared exactly i times is proportional to $if(i,k)$—that is, to the total number of occurrences of all the words that have appeared exactly i times,

II. There is a constant probability, α, that the $(k + 1)$th word will be a new word—a word that has not occurred in the first k words.

Assumptions I and II describe a stochastic process, in which the probability that a particular word will be the next one written depends on what words have been written previously. However, under these assumptions there results a reciprocity relation between words and their frequency of occurrence which is believed by Simon to apply to a quite heterogenous set of data: word counts (e.g. by Yule) where inflected forms are counted with the root word; other counts (e.g. by Zipf) where each form is regarded as a distinct word; moreover, the Yule count includes only nouns; whereas the others all parts of speech, and the Dewey, Eldridge and Thorndike counts are from composite samples compiled from a large number of different texts, whereas Yule's counts are based on a single piece of continuous prose.

Simon argues that the heterogeneity is evidence that the explanation of the reciprocity relation is to be sought in a probability mechanism, rather than in more specific properties of language. He admits, however, that the heterogeneity complicates the task of specifying the probability mechanism in detail.

This argument is fallacious. The fact that the basic assumptions to be made for constructing a mathematical model must be rather different according to the type of statistical population (e.g. whether all words or nouns only, whether from a continuous piece of text or from composite word masses) speaks decidedly against Simon's hypothesis of a common probability mechanism. In a way he admits this. If the assumptions upon which to base the mathematical structure of such a mechanism differ from one type of distribution to the other, and thus are anything but "common", how could the mechanism be so?

In fact, Simon's mathematical model is based upon a particular set of assumptions which are sensible only for a *continuous* piece of writing, and thus cannot be an adequate description of the mathematical structure of all types of count, even though

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the form of the distributions obtained by different methods of word counting were roughly similar in shape.

It follows that what the counts have in common—and it shall not be denied that they have something in common—cannot be of a purely linguistic nature. The way out of the difficulty, however, is not, as Simon thinks, to substitute a purely formal probability mechanism as the common element, which only shelves the explanation and thus fails to be one. The probability mechanism must have a material basis: something must behave according to it and thus be itself of a probability nature. In the given case we would have to say that language, if it behaves according to the probability mechanism, is the manifestation of such a mechanism, and there we are back at language as the common element. To avoid this conclusion and to regard the probability mechanism as something different from the behaviour of language, is no explanation at all, since the probability mechanism cannot exist in thin air.

In Chapter XVI we shall show that the reciprocity relation which is characteristic for the linguistic distributions examined is rooted in what is basic to all languages: the conceptual or content structure, as distinct from linguistic form or expression.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XVI

THE SEGMENTATION OF MEANING

16.1 STRUCTURE IN LANGUAGE

As mentioned in Chapter XV, and shown later in greater detail, the attempts to account for the reciprocity relation between frequency of occurrence of linguistic forms on very different planes and linguistic variables, like word length, "information", etc. cannot be regarded as fully successful from the mathematical point of view, or convincing and readily interpretable from that of the linguist. All these attempts try to deduce the relation from linguistic material, but without taking into account the conceptual basis of the use of words. We propose to show that it is the system of concepts as the totality of engrams in the brains of the members of a speech community which finds its adequate expression in the reciprocity relation of linguistic forms. In other words, the system of concepts must be regarded as the a priori structure (the "form") whose realisation ("substance") is the system of linguistic expressions.

By "structure" of concepts we understand with Hjelmslev "une entité autonome de dépendances internes", a complex whose units manifest an internal solidarity, having its own laws. Internal solidarity is here the name for the mutual relations of the concepts of a language. Be it said, once for all, that by structure in this sense we do not mean that the possible ways of connecting concepts in the course of thinking are predestined; nothing whatsoever of a "Prestabilierte Harmonie" is implied. Structure refers here only to the judicial distribution of all the objects of the manifold among the, necessarily, fewer concepts and to their mutual quantitative relations, each denoting a certain section of the universe, such sections being, of course, not restricted to the geometrical meaning of the term.

Let us imagine an assortment of articles used in the manufacture of complicated machines. There are rods and screws and nails and bolts etc. which are, of course, of such type and dimension as needed for these machines, and, insofar, the manufacture of those machines will be greatly assisted by keeping just such articles and in such dimensions in store. But that does not mean that the articles could not be used in ways different from those initially envisaged or in the manufacture of different machines than they were first intended for. Analogously with the concepts stored in the brains of the members of a speech community. They represent the store, the box of building bricks, for constructing a picture of the universe. Their "structure" lies in the storage capacity for concepts to cover all the objects of the universe—or at least

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

those objects which came under observation in the course of the development of thought—, not in the possibilities of their combination.

The most prominent and most general feature of that structure is the solidarity between them as regards content (Inhalt—definition) and size (Umfang—comprehension) of concepts. According to the general principle that the greater their size, i.e. the number of objects to be subsumed under a concept, the smaller their content, i.e. the items included in the definition of a concept, they represent a system of pigeon-holing of information to perfection.

The reciprocity principle between content and size of concepts makes it possible to reduce the space for storing the colossal amount of information, that is the colossal number of objects in the universe, in a relatively small space.

This reciprocity relation, which is already needed for the building up of the system of concepts, we regard as the a priori function ("form") whose materialisation or a posteriori appearance ("substance") is the reciprocity relation between frequency of linguistic expressions and their length.

Those familiar with Schopenhauer's epistemological ideas will not fail to notice the similarity between this conception of the reciprocity relation and his conception of causality. According to Schopenhauer, causality is already needed for the apprehension (Wahrnehmung) of the objects of the outer world, the intellect projecting the sensations of the senses according to the a priori causality function into the "outer" world; it could not therefore have been acquired by experience. Similarly, we find that a reciprocity relation is already needed for the building up of a structure of solidarity between concepts, and conclude that the reciprocity relation between symbol frequency (corresponding to size of concepts) and symbol length (corresponding to content of concepts) cannot be something peculiarly linguistic, but regard it as the a posteriori manifestation of the a priori solidarity of concepts.

16.2 CLASSIFICATION AS A STOCHASTIC PROCESS

Yule (1924) implicitly assumes that if a new species, or a new genus, arises, there is also the systematist who recognises it as such, and promptly baptises it; this does not mean that birth and baptism are simultaneous, or nearly so, but that with time every new species will be recognized as a distinct unit of the family. In short, he assumes not only a one-to-one correspondence between biological fact and name, but also, what is more important and less justified, the unavoidability of a species getting its name and place in the system, and the term as really corresponding to a definite entity in nature.

The empirical facts about systematising do not warrant these implicit assumptions. In matters of taxonomy, there arises always the vexed question of "splitters" and "lumpers". One authority asks for great differences before he separates a group of species into two genera; another only requires the slightest excuse. As a consequence, the terms "species" and "genus" may be perhaps 90% real biological entities and 10% matters of convenience or personal idiosyncrasy or, as we might say, 90% real and

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

10% imagination. When we see that there can be one series for the "lumper", and one for the man of moderate opinion, and still another for the "splitter", there can be no question of uniqueness of classification or systematisation.[1]

It is, therefore, necessary to consider the process of systematisation in taxonomy as the logical division of the "size" or extension (totum divisum) of a wider concept, into the subordinated smaller concepts (membra divisionis) according to certain principles of division, allowing always for the *penchant* of the systematist towards splitting or lumping, as the case may be. Our conception of the process is thus strictly "immanent", that is, not transgressing into the realm of phylogeny, but remaining within the range of observed facts of taxonomic work. We then see the systematist arranging his observations in certain categories which he calls "species" and "genera", whose numbers are, partly at least, at his discretion. Consequently, there is no such thing as a one-to-one correspondence between number of species and genera and discrete biological entities. Since the qualitative distinction between species, on the one hand, and between genera, on the other, is to some extent a matter of taste, it follows that the resulting numbers of either are also conditioned by the idiosyncracies of the taxonomist, his preference for either "splitting" or "lumping".

It may not be amiss to stress the difference between the distributions of genera according to the number of species per genus, and those of species according to the number of individuals per species. In the latter, the number of species among which to distribute the individuals in the sample is fixed, and so is the content of each species concept. It is thus a case of distributing a number of observations among a fixed number of defined concepts. In the former, on the other hand, it is a matter of deciding how many concepts are required for the classification of (a) individuals with related characteristics into species of a genus, and (b) the resulting species into genera of the same family. The process does not stop there, since there is no limit to the hierarchy of concepts.

Since such decisions require first a division of the *content* of concepts (partitio) into the items that go to its definition, neither the content of the concepts nor their number, be they of the first or the second order, i.e. species or genera concepts, are given. They are the result of a continued activity of the taxonomist, that is subsumption of individuals under concepts (species) and of concepts under wider concepts with less content (genera), or of the "grinding down" of content of wider species and genera into those of narrower species and genera.

For the classification process, the population of individuals must be regarded as given, and open to inspection. If we speak here of "time", it is not biological time we mean, i.e. the time of biological development of species, but "taxonomic time", i.e. the time during which classification has proceeded; in other words, the time covered by the history of biological science in the restricted sense of taxonomy.

[1] C. B. Williams, "A note on the relative size of genera . . .", *Proc. Linnean Soc., London*, Pt. 2, 1949-50.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

16.3 THE PROBABILITY MECHANISM FOR THE CREATION OF NEW CONCEPTS

I. TIME ASSUMED INFINITE

If the taxonomist is a "splitter", he will find the need for differentiation, if he is a "lumper", he won't. This is tantamount to saying that if the taxonomist is a "splitter" the probability of differentiation is p, if a "lumper" $q = 1 - p$. Since he may be either we may here speak of *chance as deciding whether or not to differentiate, that is, to introduce a name for a new species.* But we must extend the distinction between the two terms to cover mere quantitative differences between taxonomists, p being with "splitters" on the whole greater than with "lumpers", and vice versa for q. For instance, the genus Hieracium was divided by Koch into 52 species, by Fries into 106, and by Nägeli into more than 300 species. The ratio p/q indicates the extent for which the matter of classification is not chance, that is, the greater p/q, the more choice.

Let the chance of the taxonomist finding the need for introducing a name for a "new species" of the same genus, that is, the need for differentiation, in some small assigned interval of time, be p, and suppose the interval, and with it p, so small that p^2 may be ignored compared with p. Then putting aside additions to the number of genera altogether for the present, if he starts with N prime species of different genera, at the end of the interval he will have Nq genera ($q = 1 - p$) which are still monotypic, that is, have one species only, and Np genera of two species.

The need for differentiation, and with it the adding of names for new species, can now arise among the individuals belonging to the new as well as to the old species, and the matter becomes more complex. Of the Nq monotypic genera a proportion q will remain undifferentiated, giving Nq^2 new monotypes; a proportion p will be contributing Npq to the genera of 2 species. Of the Np genera already possessing 2 species at the end of the 1st interval, a proportion will remain undifferentiated, i.e. a proportion q from either species, q^2 from both species, giving Npq^2 genera of two species at the end of the second interval, or a total of $Npq\ (1 + q)$ ditypic genera altogether.

Of the same Np genera that already possessed 2 species at the end of the first interval, the proportion of either one or the other species becoming differentiated in the second interval is $2pq$, the second term in the expansion of $(q+p)^2$ giving $N.2p^2q$ genera of 3 species.

The proportion in which both of the two will be differentiated is p^2 giving Np^3 genera of 4 species. The work may be continued on the same lines, and at the end of the third interval we will have the frequencies of genera as shown in the last column of Table 68.

We must now proceed to the limit, taking the time interval t as indefinitely small, but the number of such intervals n as large, so that $n \,.\, \Delta t = t$ is finite. We may write

$$p = s \,.\, \Delta t \qquad pn = st \qquad (52)^2$$

[2] s is a constant proportional to the chance of differentiation among species occurring in a given time.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 68

Proportional frequencies of wider concepts (genera) with 1, 2, 3 narrower concepts (species) after 0, 1, 2, 3 intervals of time, the wider concept being initially monotypic (prime or primordial genera)

No. of species in the genus	Intervals of time			
	0	1	2	3
1	1	q	q^2	q^3
2	—	p	$pq\,(1+q)$	$pq^2\,(1+q+q^2)$
3	—	—	$2p^2q$	$2p^2q^2\,(1+q+q^2)$
4	—	—	p^3	$p^3q\,(1+q+6q^2+q^3)$
5	—	—	—	$2p^4q^2\,(3+2q)$
6	—	—	—	$2p^5q\,(1+3q)$
7	—	—	—	$4p^6q$
8	—	—	—	p^7

and we have the approximation

$$q^n = (1-p)^n = (1-st/n)^n \sim e^{-st} \tag{53}$$

The first term in the series of proportional frequencies is q^n or in the limit e^{-st}: that is, if f_1 is the proportional frequency of monotypic genera at time t,

$$f_1 = e^{-st} \tag{54}$$

The second term, after n intervals, is

$$pq^{n-1}(1+q+q^2+\ldots+q^{n-1}) = q^{n-1}(1-q^n)$$

or in the limit, f_2 being the proportional frequency of genera with 2 species at time t,

$$f_2 = e^{-st}\,(1-e^{-st}) \tag{55}$$

For the third term we get

$$f_3 = e^{-st}\,(1-e^{-st})^2 \tag{56}$$

The general form of the law is now obvious

$$\begin{aligned} f_1 &= e^{-st} \\ f_2 &= e^{-st}\,(1-e^{-st}) \\ f_3 &= e^{-st}\,(1-e^{-st})^2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ f_n &= e^{-st}\,(1-e^{-st})^{n-1} \end{aligned} \tag{57}$$

That is to say, if the taxonomist starts with N primary genera, all of which are monotypic, after time t the numbers of genera with 1,2,3, ... species form a geometric series with common ratio $1-e^{-st}$, s being a constant proportional to the chance of specific differentiation.

The sum of the series to infinity is unity, as it must be. As regards the mean, let us consider the general geometric series

$$S = 1 + r + r^2 + r^3 + \ldots = 1/(1-r) \tag{58}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Differentiating with regard to r, we have

$$dS/dr = (1 - r)^{-2} = 1 + 2r + 3r^2 + \ldots$$

which is the first moment about zero, so that the mean is

$$\frac{1}{(1-r)^2} \Big/ \frac{1}{(1-r)} = \frac{1}{(1-r)} \tag{59}$$

Therefore the mean of the series of f_t, say ${}_tM_s$, is given by

$${}_tM_s = e^{st} \tag{60}$$

That is to say, while the number of "splittings" and with it that of new species names within any single genus increases, of course, quite regularly according to the taxonomist's choice, the mean number within a group of genera all starting at the same time increases as an exponential function of the time.

The average number of species, that is, the size of genus, thus increases in geometric proportion: if at zero time there is 1 species, and at time 1 there are 2 species, at time 2 there are 4 species, at time 3 there will be 8 species, and so on, the number of species, on an average, doubling in successive intervals of time. This law is important, for it suggests the doubling period of species within the genus as a natural unit of time in investigations of the present kind; it is the appropriate unit of what I call "taxonomic time". In terms of years it varies, of course, from group to group.

Regarding the increase in the number of genera as the process of systematisation proceeds, it follows the same lines. If at zero "taxonomic time" there are N genera, belonging to distinct families, at time t the number of families with 1,2,3 ... genera will be given by N multiplied by the successive terms of the series

$$\begin{aligned} f_1 &= e^{-gt} \\ f_2 &= e^{-gt}(1 - e^{-gt}) \\ f_3 &= e^{-gt}(1 - e^{-gt})^2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ f_n &= e^{-gt}(1 - e^{-gt})^{n-1} \end{aligned} \tag{61}$$

where g is a constant proportional to the chance of differentiation among genera occurring in a given time. In other words, it is proportional to the probability of the taxonomist being a genera splitter, just as s was a constant proportional to the probability of his being a species splitter.

The mean number of genera in a family at time t is given by

$${}_tM_g = e^{gt} \tag{62}$$

From (60) we have

$$t = \frac{\log_e M_s}{s}$$

and from (62)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$t = \frac{\log_e M_g}{g}$$

so that

$$\log_e M_s / \log_e M_g = s/g$$

which constant we shall denote by ϱ.

The series $f_1, f_2 \ldots f_n$ gives the frequency distribution of sizes of genera for genera all of the same age t. What is the form of the frequency distribution for genera at all ages? By t we mean, of course, as always, the time of systematisation or "taxonomic time". If we start at zero time with a group of monotypic genera, both specific and generic differentiation will occur as time goes on. At time t we will actually observe, therefore, not only the frequency distribution of the primordial genera as given by the series $f_1, f_2, \ldots f_n$, as given by formula (54) to (58), but the frequency distribution of primordial and derived genera together, and this is the distribution we now require. In other words, the taxonomist's work is not restricted to differentiating the species belonging to genera of the same taxonomic age—which in most cases he would not even know, nor care to know—but his work refers also to species belonging to genera within a family which themselves have been split off from the primary genera in the process of systematisation.

We will first consider the limiting form of the distribution when t is infinite, and the primary genera become negligible in numbers as compared with the derived genera.

We find for the limiting form of the distribution, of genera of all dates of discrimination, that is for the proportional frequencies of genera according to the number of species regardless of the "age" of genera

$$\begin{aligned}
f_1 &= \frac{1}{1+\varrho} \\
f_2 &= \frac{1}{1+\varrho}\,\frac{\varrho}{1+2\varrho} \\
f_3 &= \frac{1}{1+\varrho}\,\frac{\varrho}{1+2\varrho}\,\frac{2\varrho}{1+3\varrho} \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
f_n &= \frac{1}{1+\varrho}\,\frac{\varrho}{1+2\varrho}\,\frac{2\varrho}{1+3\varrho} \cdots\cdots \frac{(n-1)\varrho}{1+n\varrho}
\end{aligned} \tag{63}$$

where ϱ is s/g. (For the derivation see Yule, 1924).

The sum of the first n terms is

$$S_n = 1 - nf_n. \tag{64}$$

The sum to infinity is unity, as it should be.

As Yule has shown, the frequency distribution (63) is one of those paradoxical distributions in which, though the median etc. are finite, the mean is infinite. This is,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of course, as it should be, for on our assumptions the mean size of a genus after infinite time must itself be infinite.

16.3.1 The bi-logarithmic straight line law

The differentiation of species from species or genera from genera seems closely parallel to the generation of offspring in a given stock if mortality is ignored; hence it seemed reasonable to assume that the proportion y of genera of age t at any given epoch would follow the law $y \propto e^{-\theta t}\, dt$. Similarly it seemed reasonable to assume that the number of species x in a genus of age t would be given by

$$x \propto e^{st}$$

Now suppose that chance can be ignored, that the number of species in a genus can be taken as a continuous variable, and that the above can be taken as absolute functional relations. The size of a genus is then absolutely determined by its age and we can find the number of genera of each size by eliminating t from the first relation. We have

$$t = \log x^{1/s}, \qquad dt = (sx)^{-1}\, dx$$

whence

$$\begin{aligned} y &\propto x^{-(g+s)/s}\, dx \\ &\propto x^{-(1+\rho^{-1})}\, dx \end{aligned}$$

This suggests the bilogarithmic straight line relation $yx^k = C$. A closer study of the bilogarithmic plots of the data in question shows, however, that this rather crude approach does not work. Up to genera of 30 species or so, one could speak of a straight line relationship, but after that the points begin to drop away from the line, i.e. the observed frequencies of the larger genera were smaller than those suggested by the straight line law. The fundamental mistake in assuming the straight line law relation between the logarithm of the observations lies in the assumption that time might be regarded for practical purposes as infinite: it must, in fact, be regarded as quite short, which speaks in favour of the interpretation of time as that of systemisation, and not as that of biological development.

16.4 THE PROBABILITY MECHANISM OF CREATING NEW CONCEPTS (CONTD.)

II. TIME ASSUMED FINITE

The limiting form of the frequency distribution when time is infinite, as derived in the last section, must be corrected if it is to fit series for which time is finite. The correction factors c_t, the subscript agreeing with that of the proportional frequency f_t to which the correction applies, whose derivation is given in Yule's paper, are

$$c_1 = c_1'$$

$$c_2 = c_1' - c_2'$$

डॉ० विद्या निवास मिश्र
प्राध्यापक, भाषा विज्ञान
निजी पुस्तकालय
पुस्तक संख्या.....

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$c_3 = c_1' - 2c_2' + c_3'$$
$$c_4 = c_1' - 3c_2' + 3c_3' - c_4' \qquad (65)$$
. . . . . .

and so on, where one new term c' is introduced at each stage, and the numerical coefficients are the coefficients of $(1-1)^{n-1}$, that is, the binomial coefficients, the terms c_i' being

$$c_1' = \varrho\,(1+\varrho)^{-1}\, e^{-\alpha\varrho^{-1}(1+\varrho)\tau}$$
$$c_2' = 2\varrho\,(1+2\varrho)^{-1}\, e^{-\alpha\varrho^{-1}(1+2\varrho)\tau}$$
$$c_3' = 3\varrho\,(1+3\varrho)^{-1}\, e^{-\alpha\varrho^{-1}(1+3\varrho)\tau} \qquad (66)$$
. . . . . .

where $\alpha = \log_e 2 = 0.6931$ and τ is the number of doubling-periods, that is the number of times in which the number of species within the genus is doubled. The frequencies of species at time τ are then given by

$$f_1 = f_1 + c_1$$
$$f_2 = f_2 + c_2$$
$$f_3 = f_3 + c_3 \qquad (67)$$
. . . . . .

For fitting the theoretical distribution to actual frequency data for size of genus, we shall follow the method suggested by Yule. This method was to fix on the proportion of monotypes and the mean number of species to the genus as the characteristics to be used for determining ϱ and τ; for the determination of ϱ and τ to use a table giving the proportion of monotypes (f_1) and the mean number of species to the genus (M) for all values of ϱ and τ that seemed likely to occur, and in a given case to determine them from this table by inverse interpolation (Appendix F).

16.5 ILLUSTRATION

For fitting the theoretical distribution of content, we have chosen two types of empirical data: zoological and linguistical. The first is the distribution of genera of snakes according to the number of species, the other is the distribution of sub-columns in the Chinese dictionary according to column length, that is the number of characters per column (see Vol. I, Sect. 13.2.1 to 13.2.5). This linguistic distribution is fully comparable to that of zoological data. Every character stands for a species. A radical plus an n-phonetic ($n = 1, 2, 3 \ldots$ strokes) stands for genus. The distribution of n-phonetic subgroups to a radical according to the number of characters in the subgroup thus corresponds to the distribution of genera according to that of species per genus. It is thus a distribution of wider concepts—in the linguistic, not content sense

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 69

Snakes: The Observed and Calculated Numbers of Genera of each size

Number of Species in Genus	Number of Genera	
	Observed	Calculated
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upwards	4	6.2
Total	293	293.0

TABLE 70

Chinese Dictionary

x	y	
	Observed	Calculated
1	714	710
2	332	290
3	174	171
4	113	117
5	82	86.8
6	59	67.7
7	44	54.5
8	41	45
9	27	37
10	29	31
11	19	27
12	19	23
13	15	19.9
14	20	17
15	17	14.6
16 to 20	49	44
21 to 25	28	32.8
26 to 30	17	24
31 to 40	26	13
Total	1825	1825.3

x = length of sub-columns to radicals in a Chinese Dictionary (Mathews) (i.e. no. of combinations consisting of a radical and different phonetics having the same stroke number).

y = frequencies of sub-columns of the same length.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

—according to the smaller concepts comprised by them, and thus should conform with the theoretical distribution, provided the latter truly fitted the distribution of content.

The method of fitting consists, as explained above, in using the proportion of monotypes and the mean number of species per genus as the characteristics for determining ϱ and τ; a numerical table from which to read off these values for any pair of proportion of monotypes and mean number of species per genus is given in Appendix F. That table gives f_1 and N for all values of ϱ from 1.0 to 3.0 by 10ths of a unit, and for all values of τ from 1 to 10 by units. The exact levels of both these parameters have, as a rule, to be found by interpolation. We list below particulars respecting both the zoological, and the linguistic distribution. The values of χ^2 show the theoretical distribution as fitting the observed data satisfactorily.

TABLE 71

	Snakes	Chinese Dictionary
Number of Genera	293	1825
Number of Species	1475	8765
Mean species per genus, M	5.034	4.803
Proportion of monotypes, f_1	0.4471	0.3913
τ	4.260	3.500
ϱ	1.253	1.720
χ^2	13.77	22.125
n (Degrees of Freedom)	11	17
P	0.18	0.18

16.6 HOW CONCEPT CONTENT CHANGES WITH CONCEPT SIZE

The frequency distribution of genera according to the number of species per genus gives information about "concepts" in contradistinction to "expression" (we had better avoid the term "content" for "concepts" because it is the recognised term used in logic for the items in the definition of a concept, its "intension" as different from its "size" or "extension"), but directly only about the extension aspect of "concept", its "size" aspect. It tells us how all the species, and thus the individuals comprised by them, distribute themselves among the genera. Indirectly, of course, we can infer from the size of genera their content, since in general the two are inversely related, but directly the distribution tells us nothing about what happens to the content of a concept if its size increases or decreases.

Since this is just what would characterise concept structure, or structure of meaning, it seems to be worthwhile to go into the matter and see whether characteristic para-

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

meters of the structure of meaning in this sense could be obtained. Our argument leading up to such a result is as follows. As concept size increases, its content decreases. Processes of gradual and continued comminution of size of physical objects have been studied and treated mathematically, the best known being that of grinding solid materials. Using the most general mathematical function for the comminution process, we shall apply it to the comminution of concept content with increase in size. We do not precisely know the number of items in the definition of each concept, but we do know that it will be inversely proportional to the concept size in terms of number of individuals or of lesser concepts subsumed under it. We shall, therefore, fit the mathematical function for grinding to our distribution of genera, using the *reciprocal* genus size as variable. If we succeed in describing in this way the comminution process of content, as wider concepts are derived from narrower ones, it will provide us with the characteristic parameter we are looking for, and prove that meaning is structured in this sense.

16.7 THE ROSIN-RAMMLER LAW APPLIED TO THE REDUCTION OF CONTENT WITH INCREASE OF SIZE OF CONCEPTS

Let us suppose that the distribution of a sample of particulate matter is obtained by sieving. Calling the quantity (in %) which passes the sieve, that is the weight percentage of particles smaller than the sieve opening, D, and the quantity retained on the sieve, that is the percentage by weight of particles bigger than the sieve opening, R, we obtain by plotting either D or R against the particle size, the fineness characteristic curve, D and R being connected for every sieve opening by the relation $D + R = 100\%$.

From probability considerations, one deduces that the weight distribution curve is governed by the law

$$y = \frac{dG(x)}{dx} = 100\, nbx^{n-1}\, e^{-bx^n} \tag{68}$$

where x is the particle size in terms of the diameter of (assumed) spherical particles, n and b are constants, b increasing with increasing fineness, and thus being comparable to a measure of spread, and n being independent of fineness but characteristic of the substance.

Integrating, we get

$$D = G(x) = 100\,(1 - e^{-bx})^n \tag{69}$$

for the undersize weight distribution, and

$$R = 100 - G(x) = 100\, e^{-bx^n} \tag{70}$$

for the residual weight distribution.

Taking logarithms twice, we have

$$\begin{aligned} \log R/100 &= -bx^n \log e \\ \log(-\log R/100) &= \log b + \log\log e + n \log x \\ \log\log 100/R &= \mathrm{C} + n \log x \end{aligned} \tag{71}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

This represents a straight line relationship between log log $100/R$ and log x. If the series of corresponding values of these quantities plots sensibly to a straight line, it would support the view of the Rosin-Rammler curve as a satisfactory description of the grinding process.

If the constant n as the slope of the straight line is obtained geometrically, equation (71) can be solved for b. Having numerical values for both n and b, R can be calculated by (70) for any specified value of x; n is an index of the range of sizes, and b of the degree of comminution.

For practical purposes a graphical method is available based upon the Rosin-Rammler law, which saves the trouble of the double logarithmic transformation of $100/R$ and of the logarithmic transformation of x; graph paper whose grid is such as to have these variables already in log log respectively log transformation is obtainable commercially.

Using the graphical method for our purpose of analysing the diminution of concept content, we must remember that content is inversely proportional to size of concept

$$\text{content} \propto \frac{1}{\text{size}} \tag{72}$$

Since the factor of proportionality—which we do not know—is immaterial for the relation between frequency of concepts and content, we plot on the horizontal ordinate of the grid the reciprocal of the genus size; on the vertical ordinate we plot, of course, the proportional frequency of genera of specified *size or less*, which corresponds to specified *content or more*.

In order to obtain a parameter characterising the degree of comminution of concept content through continued differentiation of families into genera and of genera into species, we use the following considerations.

The constant b is an index of the degree of comminution of content, such that its reciprocal $1/b = \mu$ is the content corresponding to the maximum or peak of the distribution curve. For $n = 1$ (a line inclined at 45°) and $x = \mu = 1/b$, formula (70) becomes

$$R = 100/e = 36.8\% \tag{73}$$

This means that the point where the straight line on a Rosin-Rammler grid crosses the horizontal of 36.8%, projected on to the horizontal axis represents the degree of content for which the distribution curve has a maximum, or as we say, its peak. This suggests the use of n/μ as a characteristic constant of the distribution of comminution of concept content.

Illustration. The two series of data to which before the theoretical distribution of genera according to species was fitted are now used for applying the Rosin-Rammler law and especially the method of graphically fitting a straight line to the data according to that law (Tables 72, 73, and Fig. 13).

We find a sensibly linear relation in accordance with the assumed mathematical

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

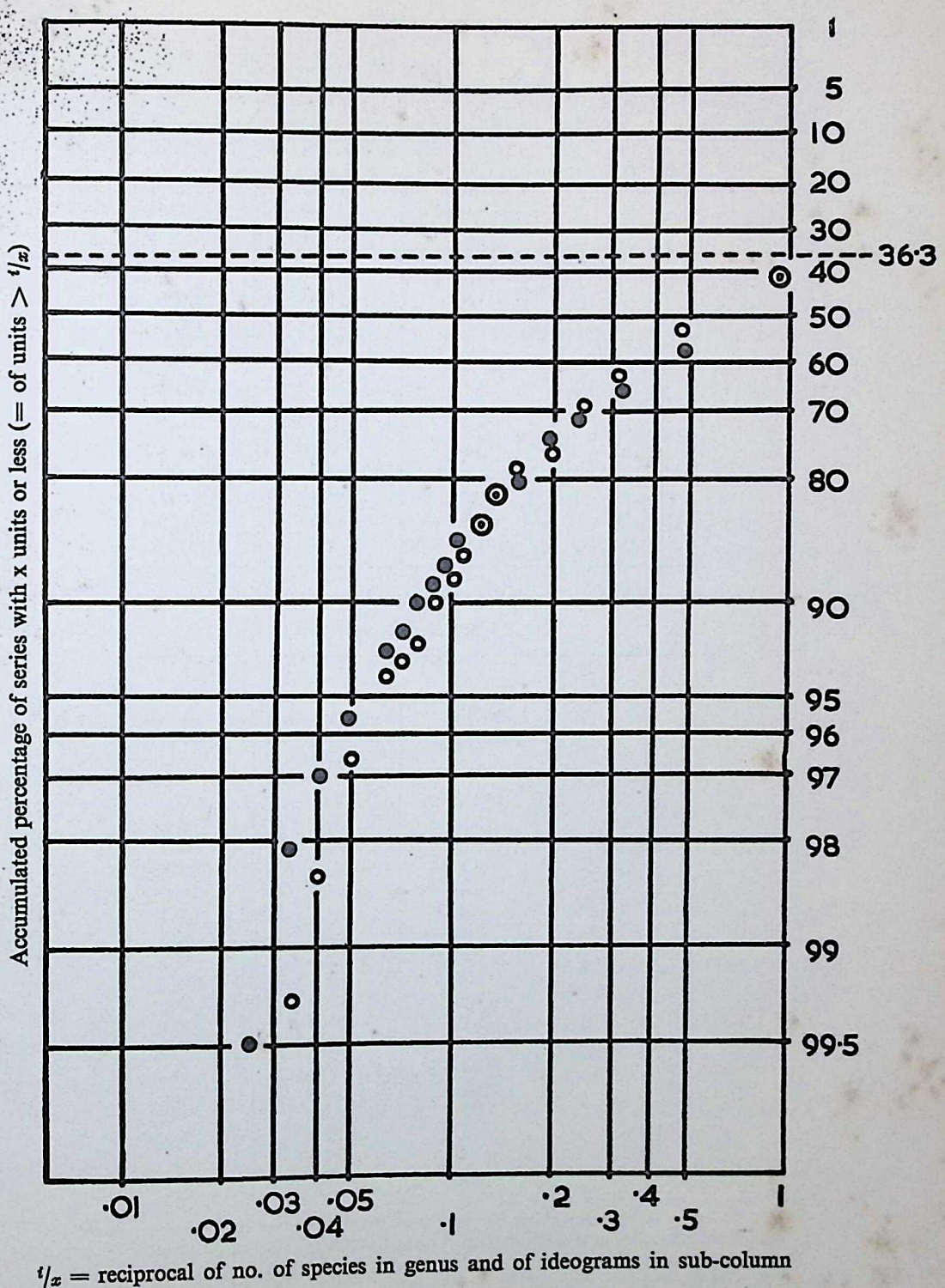


FIG. 13. Accumulated percentage of genera with x species [○], or of sub-columns with x ideograms [●], or less, plotted against the reciprocal $^1/_x$.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 72

Snakes: The Observed and Calculated Numbers of Genera of each size

Number of species in genus	Number of genera		
	Relative Frequency (%)	Cumulative Relative Frequency (%)	
		Observed	Theoretical
1	44.7	44.7	44.7
2	11.9	56.6	60.8
3	9.6	66.2	69.3
4	5.8	72.0	74.8
5	5.5	77.0	78.6
6	3.1	80.6	81.5
7	2.7	83.3	83.6
8	2.7	86.0	85.2
9 to 11	4.4	90.4	89.2
12 to 14	1.0	91.4	91.6
15 to 20	2.4	93.8	94.4
21 to 34	4.8	98.6	97.7
35 upwards	1.4	100	100

TABLE 73

Chinese Dictionary

x	y		
	Relative Frequency (%)	Cumulative Relative Frequency (%)	
		Observed	Theoretical
1	39	39	38.9
2	18.2	57.2	54.8
3	9.5	66.7	64.2
4	6.2	72.9	70.9
5	4.4	77.3	75.7
6	3.2	80.5	79.4
7	2.4	82.9	82.4
8	7.2	85.1	84.9
9	1.4	86.5	86.9
10	1.5	88.0	88.6
11	1.0	89.0	90.1
12	1.0	90.0	91.4
13	0.8	90.8	92.5
14	1.1	91.9	93.4
15	0.9	92.8	94.2
16 to 20	2.7	95.5	96.6
21 to 25	1.5	97.0	98.4
26 to 30	1.1	98.1	99.7
31 to 40	1.4	99.5	100.4

x = length of sub-columns to radicals in a Chinese Dictionary (Mathews). (i.e. no. of combinations consisting of a radical and different phonetics having the same stroke number).

y = frequencies of sub-columns of the same length.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

structure of the Rosin-Rammler law. For both examples, the distribution of genera of snakes according to reciprocal size, and of Chinese dictionary sub-columns according to reciprocal length, the curve has its peak close to the theoretical maximum, as it should be, since they are both inversely *J*-shaped. That the line bends towards the finer contents is something which it has characteristically in common with the experimental results of fine grinding of minerals, and lends support to the conclusion that the mathematical structure of comminution of concept content with time is similar to that for particulate material with time of grinding.

16.8 ASPECTS OF QUANTITATIVE LANGUAGE STRUCTURE

The aspect of "structure" discussed in the preceding section was that of "storage", that is, the peculiar nesting of concepts according to content and size which enables us to subsume every observation under an appropriate concept and denote it by an appropriate word. Such an arrangement of ideas is, as Samuel Butler expressed it, "as much a matter of convenience as the packing of goods in a druggist's or draper's store". And he adds: "Truth is only that which is most largely and permanently time-saving." This applies even to mathematical theories; as Poincaré has shown; what is required of a geometry is not to be true but to be convenient.

Another aspect of structure, closely connected with that of solidarity of size and content of concepts which, in the plane of language expression, becomes that of frequency and meaning, is that of the method for achieving distinction between linguistic symbols. Adopting Trubetzkoy's principle of the phonological opposition as the principle of distinction, we arrive again at the idea of a solidarity among linguistic forms as part of linguistic structure. Since it is not the particular phonemes, but the phonological oppositions implied by the distinctive use of phonemes which is the basic notion, it is evident that we are again faced with a solidarity among all possible oppositions, as far as they are admissible in the language under consideration.

The third property of language structure, with which we shall deal in Part IV, is that of duality, one aspect of which is the solidarity of content (signifié) and expression (signifiant). If language is to be what it should be, there must be no one-sidedness or preference between content and expression. Starting from the one, it must be possible to arrive at the other, and any change in one must be accompanied by a corresponding change in the other. In other words, and somewhat more precisely: it is an essential of language as a system of reference for the events of the manifold that either the plane of content or that of expression should be suitable as a system of reference. Thinking in terms of content of linguistic forms leads to a certain complex of linguistic expressions, and any change in the complex of thought is accompanied by a corresponding one in the plane of expression. And conversely, starting with a complex of expression, there must be a corresponding complex of thought in the plane of content, and any change in the former entails a corresponding change in the latter.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

FOURTH PART

LINGUISTIC DUALITY

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XVII

THE TYPE-TOKEN DUALITY

Information theory which conceives of language as a system of coding supposes that the message which is to be transmitted in the form of linguistic symbols must exist in the sender's mind before being coded. That the message when de-coded at the receiving end of the communication channel should be precisely that encoded and sent, is regarded as the Acme of perfection of the coding-plus-transmission system. Language as a creative force, as an art, has no place in information theory. It is, therefore, small wonder that many literary people fail to see what relation that theory has to language as they understand it. The conception and description of language as a system of coding, however ingenious, they regard as little more than commonplace, since it turns a blind eye to the more recondite functions of language as they understand the term.

The case is reminiscent of the two ways of interpreting music. Music, always valued as an art, came, in more recent times, to be described by Bernard Shaw and others as a code for the expression of emotion. Bernard Shaw, for instance, draws a sharp line between the music of Mozart and Haydn, and 18th-century composers in general, on the one hand, and that of Beethoven, on the other: "The music of the 18th century is all dance music: a symmetrical pattern of sound that is pleasant to move to. Music, however, can do more than make beautiful sound patterns. It can express emotion. Now all Beethoven did was that he used music altogether as a means of expressing emotion."

He did not notice that what he thus gave out to be the last word about music—this aspect of music had really been much more profoundly treated by Schopenhauer almost half a century before Shaw put forward his theory—was little short of a platitude, since nobody had ever doubted that music had something to do with emotion. What Shaw so disparagingly calls the beautiful sound pattern, which is perhaps best exemplified by the Fugue, reveals the property which is essential to musical creation, viz. that music has its own laws which are conditioned, not by extra-musical thought and emotion, but by the very material of music itself. Composing in such a way as to let these laws reign supreme is what makes the composition a work of art. Only in this sense can music be regarded as creative.

We shall show that language, apart from being a medium or "channel" for express-

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

ing thought, or, as we say, for communication, is governed by an intrinsically linguistical principle of duality, such that if a statement is made it provides the occasion for another statement, and so on. Language is thus self-generating, every statement being possibly productive of another. *Language not only expresses thought, it also creates thought.* This aspect of language was emphasised in an inimitable way by the Viennese writer K. Kraus. In this sense, we may say that it is duality which makes language tick.

Considering that the law of duality is the fundamental law of language, it is clear that duality must permeate all linguistic work, though to a different extent. The more linguistic expression is meant to serve the purpose of communication, the less duality will be found in it, and vice versa. This is true, no matter whether duality is realised to be a factor of language, or whether it is applied without such express realisation.

17.1 THE PRINCIPLE OF GEOMETRICAL DUALITY IN LANGUAGE—INTERCHANGEABILITY OF TYPE AND TOKEN IN LINGUISTIC STATEMENTS[1]

Geometrical methods have been used in structural linguistics so far only tentatively, not systematically. Nor was the question ever asked whether such methods as were used really fitted language data satisfactorily, and if they did, why they should do so. A more fundamental approach is needed, and in what follows we shall make a step quite analogous to the one by which the theory of statistics was made applicable to language through equating the total of word engrams ("la langue" in de Saussure's terminology) plus their probability of occurrence to the concept of the statistical population, and the individual speech utterances ("la parole") to random samples from such a hypothetical population (Vol. I, Sect. 5.3).

Into any set of axioms there must enter certain undefined concepts, such as "point" and "line" in geometry. They can be regarded as purely abstract entities whose properties in a deductive system are given entirely by the relations that hold between them as stated by the axioms. The totality of axioms provides the *implicit definition* of all "undefined" geometrical concepts, such as "point", "line", "incidence", etc. For applications it is important that the concepts and axioms of geometry correspond well to physically verifiable statements about "real" tangible objects. Thus, even for the formalist there is an authority other than the human mind that decides the direction of mathematical thought. Einstein has expressed this as follows: "Euclidian geometry, considered as a mathematical system, is a mere play with empty concepts; straight lines, planes, points etc. are mere 'fancies'. If, however, one adds that the straight line is to be replaced by a rigid rod, geometry is transformed into a physical theory."

The physical reality behind the concept of "point" is that of a very small object, such as a pencil dot, while a "straight line" is an abstraction from a stretched thread or a

[1] The Epilogue (III) should be read in conjunction with this and the following section.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

ray of light. The properties of these physical points and straight lines are found by experience to agree more or less with the formal axioms of geometry. Quite conceivably, more precise experiments might necessitate modification of these axioms if they are to describe adequately physical phenomena.

Following this method we now let a "point" correspond to a linguistic form (phoneme, word, morpheme, tagmeme etc.) which we shall briefly refer to as TYPE, and "line" to the linear sequence of such forms (of phonemes in words, of morphemes and words in sentences, of tagmemes in grammatical combinations, etc.) which we shall refer to as TOKEN. We need not worry about these concepts not being strictly defined; this is what they have in common with their prototypes "point" and "line", and the implicit definitions will emerge from the totality of their relations in language.

The above remarks about the required agreement in properties between geometrical concepts and their physical counterparts apply, mutatis mutandis, here too. Experience must show whether the properties of the linguistic "point" and "line", i.e. type and token, agree more or less with the formal axioms of geometry. And very likely, modifications may be necessary in the formal axioms if they are to describe adequately linguistic phenomena. Only if the formal axioms agree more or less with the properties of linguistic objects will geometry be of interest to the linguist. In other words, the statements of geometry must remain true in terms of type and token.

As a criterion for the correctness of our claim to have made geometry truly applicable to language structure by letting "type" stand for "point", and "token" for "line", we use the property of DUALITY, *according to which the theorems of projective geometry remain true if "point" and "line" are interchanged,* and thus occur in pairs, each similar to the other, and, so to speak, identical in structure. This relationship is called duality, and gives projective geometry a character quite distinct from that of elementary (metric) geometry, in which no such duality exists. In the plane this depends upon the fact that just as two points determine a line in which these points lie, so two lines determine a point which lies on both the lines. Thus, if we have any plane figure consisting of points and lines, there exists also a figure consisting of lines and points such that to any point of the first figure lying on a line of that figure there corresponds a line of the second figure passing through a point of this figure which corresponds to the line of the first figure, and so on.

Two such figures as we have described are said to be *duals* of one another. They may be the same figure which is then said to be *dual to itself.* For instance, a triangle; it is a simple affair of three points; but it also has three lines, the points constituting its vertices and the lines its sides. Just as two vertices lie on one of these lines, two sides pass through one of its points, and any such single statement cast into this form can be balanced by another statement, itself dual, wherein the words "line" and "point" interchange their roles. The example is absurdly simple, but nonetheless contains the germ of a vast and wonderful system that unfolds itself from one dimension to another.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

17.2 DUALITY OF LINGUISTIC THEOREMS

The principle of duality in a plane asserts that if in any valid proposition of plane projective geometry the words "point" and "line" are interchanged, the resulting proposition is also valid. My contention is that an analogous principle applies to language, and that many controversial questions arise out of insufficient realisation of that astounding phenomenon.

The principle of linguistic duality asserts that if in any valid proposition of structural linguistics the words TYPE (linguistic form, e.g. phoneme, morpheme, word; content of concepts) and TOKEN (frequency of occurrence and probability of linguistic forms; number of cognitions) are interchanged, the resulting proposition is also valid.

For instance, it was stressed in Sect. 12.3 Vol. I, that the vocabulary-occurrence relationship represented the fundamental linguistic duality. There are two distributions involved: the individual items of vocabulary can be listed on the abscissa as the variable, and the frequency of each is plotted as the ordinate (Zipf, 1929); but the same data can be differently plotted by reversing the roles of vocabulary and frequency. We could list on the abscissa the various frequencies that occur, and plot on the ordinates the number of vocabulary items showing specified frequencies (Yule, 1944). The two plots are not contradictory or competing descriptions; rather they are complementary ways of summarising the same data.

Dualities are preferably set out in twin columns, like translations of one another.

The fundamental linguistic duality (Sect. 12.3.1, Vol. I)

The TYPE-TOKEN relation can be presented by

the distribution of frequency of occurrence (token) with regard to the individual items (type) of vocabulary.	the distribution of vocabulary items (type) with regard to their frequency of occurrence (token).

The "Characteristic" K duality (Sect. 5.3)

In Sect. 5.3 it was shown that according to which distribution one has in mind, that of word occurrences according to the individual words, or that of words according to their frequency of occurrence, the characteristic K is derived either as a repeat rate of individual words or as the coefficient of variation (squared) of the mean frequency of occurrence of words.

In twin columns we have, for a particular linguistic distribution or a given literary text:

the sum of squared relative frequencies (token)—repeat rate—of words (type)	the squared relative variation of the distribution of words (type) according to frequency (token)

is a constant "K".

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The fundamental theorem of information theory and the $K \leqslant$ antilog $(-H)$-duality (Sect. 14.2)

The expectation of a *linguistic form* (e.g. word length) is an upper limit to the negative expectation of the *log* of the probability of the linguistic form:

$$M \geqslant H$$

the expectation of the *probability* of a linguistic form is an upper limit to the *antilog* of the negative expectation of log probability of the linguistic form:

$$K = E(p) \geqslant \text{antilog}\,(-H) = -E(\log p)$$

The basic relations of TYPE-TOKEN mathematics and of statistical linguistics are thus seen to follow the patterns of linguistic duality as defined above. Other dualities are:

The Phoneme and Phoneme Distribution Duality (Sect. 8.3) (Alphabet and Alphabet Distribution Duality)

The phonemes of a language are the *type* whose use as *token* in building up the words of the language determines their functional burdening.

The phoneme occurrence in speech and writing is the *token* of the distribution of phonemes in the dictionary, their functional burdening, as the *type*.

The Word Length Duality (Sect. 9.2)

The distribution of words in a text according to length in terms of phonemes, letters or syllables is the *token* of the corresponding distribution of word length in the dictionary as the *type*.

The distribution of the different vocabulary items in a text according to length in terms of phonemes, letters or syllables is the *type* of which the corresponding distribution of word length occurrence in the text is the *token* (Moment distribution property of word length as a log-normal variate).

The Dictionary and Text Duality (Sect. 2.1)

The vocabulary items of the dictionary are the *type* whose repeated use in texts as *token* results in the frequency distribution of words.

The distribution of vocabulary items according to frequency in a text is the *type* of which the distribution of occurrence numbers in that text is the *token* (Moment distribution property of word counts as log-normal variates).

17.3 LINGUISTIC DUALITY AND UNI-LATERALITY OF THE SPEECH FUNCTION IN THE BRAIN

A consequence of the principle of duality is that *for each theorem that is proved another theorem is obtained, and the latter is true and does not require a new, direct proof. The number of theorems is thus automatically doubled.* The word "automatically" should not be understood too literally: although there are instances which appear almost

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

self-evident, and thus are not very exciting, yet there are many dualities which are anything but obvious and which may be regarded as the very foundations upon which the structure of projective geometry rests, such as Pappus' theorem and Pascal's and Brianchon's theorems.

This doubling of the number of theorems as a consequence of duality is of the greatest importance in language, where it renders redundant a duplication of linguistic events, such as could be achieved by bi-laterality of the corresponding brain function. This accounts for the strange fact of the uni-laterality of the most complicated brain function: speech.

Whereas the sensations of sight, hearing, touch, balance and those connected with the motor activity of the arms and legs occur in symmetrical pairs due to the bi-laterality of the corresponding brain functions, speech is located in one lobe only, the left, and is thus uni-lateral. *It can dispense with bi-laterality because it is intrinsically such as to provide a double, the dual, to each of its relations.* This compensates fully for possible disadvantages which may be attached to uni-laterality. The automatic shift of viewpoint brought about by the duality mechanism does for linguistic observations what bi-laterality of, say, sight does for visual observations.

The astounding discovery of geometrical duality led to a strange conclusion, namely that the nature of the basic entities to which the postulates of a deductive science are applied is quite immaterial. "The importance of the 'windfall' which the principle of duality meant for geometry is quite obvious. But the philosophical by-product which that principle entails is even more far-reaching. We have a theorem dealing with certain entities, namely points and lines. If this theorem remains valid when these entities are replaced by others, namely lines and points respectively, then our original theorem is not specifically a statement about points and lines."[2] Applying this to language, we come to the conclusion that what is invariant in any statement about Types and Tokens is the relation between these entities, not the description of parts of the manifold or of our consciousness as either Types or Tokens (Epilogue III).

17.3.1 Bi-lateral and uni-lateral brain functions[3]

In speech a uni-lateral mechanism of the brain is employed for the first time, as opposed to the bi-lateral mechanisms of articulation and vision.

Both hemispheres share in the acquisition and performance of speech. How comes it that in the average individual it is only necessary to destroy a limited portion of the left cerebral cortex for a complete loss of speech to result? One explanation offered is the general right-handedness of homo sapiens, but the inadequacy of this solution can be seen when it is realised that the difference between the daily work of the two arms is vastly less than the corresponding difference in function between the right and left speech areas. The purely mechanical explanation that right-handedness stimulates the left hemisphere would imply that writing was prior to speaking, and thus to speech

[2] N. A. Court, 1956.
[3] H. Jackson, 1864; H. Head, 1920; S. E. Henschen, 1920; S. A. Kinner Wilson, 1926.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

and language in general, which reverses the order and importance of the two aspects of language: the spoken and the written. For all Western languages at least, or better for all languages except some Far Eastern ones, speech or sound is the primary element and writing the secondary element of language.

Excepting language, we must regard the differentiating activity of both hemispheres of the brain to be the rule. Specified visual sensations in both eyes usually accompany those sensations of touch from which we derive the ideas of size, distance, form. The area of the retina, which is occupied by the picture of a big object, is, conditions being equal, greater than that occupied by that from a small object; the area covered by the picture of a near object is greater than that covered by the picture of an equally great but remoter object; conversely, if these sensations arise, they evoke inevitably the image of objects which would produce the corresponding touch sensations.

If an object is recognised by touch as being one, the centres of its picture on both retinas fall into the centre of the yellow spot, if both eyes are directed towards the object; if, on the other hand, there are two objects instead of one, the centres of their two retina pictures cannot fall simultaneously upon the centres of the yellow spot in both eyes. Conversely, if the centres of two pictures simultaneously reproduced in both eyes fall in the centres of the yellow spot, we conclude that there is only one object responsible for it. If they do not coincide with the centres of the yellow spot, then we conclude that two objects are responsible for the sensations.

If we see with both eyes, the axes of both eyes, of whose movement the muscle sensation informs us, cross each other in a greater angle if the object is near, and in a smaller angle the more remote it is from us. Conversely, if by the unchanged position of an object, the axes of the eyes directed towards it cross at a greater or smaller angle, we conclude that the object is nearer or further away.

Language is altogether different in this respect. Whereas in the sensory and motor activities of the body a bi-lateral brain mechanism is necessary, if not for performance altogether, so yet for differentiation, control and efficiency, the structure of language is such as to make a bi-lateral differentiation unnecessary, because the element of duality is inherent in language.

17.3.2 The doubling mechanism of language

Linguistic phenomena may be regarded either from "without" or from "within", either from the outward form (the *token* in the sense of symbol) or from the inner content (the *type* in the sense of function or meaning).

In the first case we take the sound (of a word, or some other part of a linguistic expression) and inquire into the meaning attached to it; in the second case we start from the signification, and ask ourselves what formal expression it has found in the language we are dealing with. If we denote the outward form, the token, by the letter 0, and the inner meaning, the type, by the letter I, we may represent the two ways as $0 \rightarrow I$ and $I \rightarrow 0$, respectively. We shall now show this doubling mechanism as permeating language in its lexical and grammatical aspects.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

1. A one-language dictionary may be arranged in different ways. Starting from without ($0 \rightarrow I$), we take a word and then explain what it means. The various meanings of the same word are given, and in some instances these may, in course of time, have become so far differentiated as to constitute different words. This leads to placing together words that have the same sound (homophones or homonyms).

Or we may start from within ($I \rightarrow 0$), and try to arrange all the things and relations signified in a systematic or logical order. (An attempt was made in Roget's *Thesaurus of English Words and Phrases*.) Here all synonyms are placed together. Because of the difficulty of a truly objective and consistently systematic arrangement of all special facts, most dictionaries content themselves with an arrangement in alphabetical order as a convenient, though unscientific, substitute for the $I \rightarrow 0$ type of arrangement.

Substituting "type" for meaning or content, and "token" for sound (word form), the two methods and the resulting arrangements

$$\text{Token} \rightarrow \text{Type} \quad (0 \rightarrow I)$$

and

$$\text{Type} \rightarrow \text{Token} \quad (I \rightarrow 0)$$

are seen to be dual to one another.

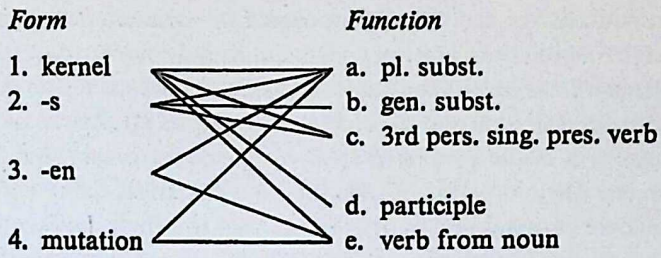
2. In grammar, too, we may start from without or within. In the first case ($0 \rightarrow I$), we take a form and inquire into its meaning or function; in the second ($I \rightarrow 0$), we start with the meaning or function and ask how that is expressed in sound.

The conventional paradigmatic method gives in one place all the forms of the same word, like Latin servus, serve, servum, servo; amo, amas, amat, amamus, etc. Thus the paradigmatic arrangement is not one according to grammatical form, since it brings together different forms of the same word which only belong together from a lexical point of view. The new system of division of grammar may start from sound, and thus from without ($0 \rightarrow I$). For instance, in speaking of the ending -s in English, we mention first its function as a sign of the plural in substantives, then as a genitive sign, then as a mark of the third person singular in the present tense of verbs, then in the non-adjunct form of possessive pronouns, e.g. "ours". The ending -n (-en) in English serves in a similar way to form a plural in "oxen", a non-adjunct possessive in "mine", a participle in "beaten", an adjective in "silken", a derived verb in "weaken", etc. The second main division of grammar starts from within ($I \rightarrow 0$). The facts of grammar are the same in the two methods of subdivision, but the point of view is different, the treatment is different, and the two methods of classification supplement each other and together give a complete and perspicuous survey of the general facts of grammar. To show that the double method of approach leaves us with a clearer picture of the intricate network of such a language as English than would have been possible to achieve by the older paradigmatic method, we tabulate below one part of this network exhibiting the manifold cross-strains of form and function.[4]

[4] O. Jespersen, 1924.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri



Examples:

1a... sheep;	1c... can;	1d... put;	1e... hand;
2a... cats;	2b... John's;	2c... eats;	
3a... oxen;	3e... frighten;		
4a... feet;	4e... feed.		

Substituting "type" for function and "token" for form, the two ways of arranging grammar according to whether we start from within or without, namely

$$\text{Type} \rightarrow \text{Token} \quad (I \rightarrow 0)$$

and

$$\text{Token} \rightarrow \text{Type} \quad (0 \rightarrow I)$$

are again seen to be dual to one another.

It is now important to realise that the two ways of arranging dictionary items or grammar forms are not exclusive to one another, but, as duals, are complementary and, consequently, are used one with the other. This seems to some extent inevitable. The diligent use of our conventional one-language dictionaries must lead, by and by, to the material arranging itself in our heads according to meaning, however great the differences in this respect may be between different individuals. The matter is more obvious in the case of language translation. We may get used—and often do so—to working with a one-way dictionary only, say English-German, when translating, no matter which language we take as the source and which as the target language. But as time goes on, the matter will arrange itself in our heads, to some extent, also according to German as the source and English as the target language. The reason is that there is no such thing as a one-way translation in the strict sense of the term. We cannot really translate language *A* into language *B* without mentally performing also the opposite transaction, that is without being constantly, though less distinctly, aware of the *A*-equivalents of given *B*-words.

Comparing the two arrangements of grammar forms, according to function (type) and according to sound (token), we discover that the two viewpoints are those of speaker (writer) and hearer (reader), respectively. The speaker starts from certain ideas; to him the meaning or function is the given thing, and he has to find out how to express it; he moves from within to without, $I \rightarrow 0$, from type to token. The hearer,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

on the other hand, as the recipient of information, encounters certain sounds and forms and has to find out their meaning or function; he moves from without to within, $0 \rightarrow I$, from token to type. Again, these methods are not exclusive of one another, quite apart from the fact that the speaker is not always a speaker and the hearer not always a hearer. As in the case of the one-language dictionary and in that of the one-way bi-lingual dictionary, the hearer could not effectively distinguish between the tokens and quickly recognise them, that is use the $0 \rightarrow I$ method, without the $I \rightarrow 0$ method being, to some extent, also present in his mind; and the speaker could not, for any length of time, select the correct form quickly, that is use the $I \rightarrow 0$ method, without the $0 \rightarrow I$ method taking firm root in his mind.

Grammatical forms, like dictionary items, are Janus-like: they face both ways towards "type" and "token", and the result is the duplication, in the sense of duality, of the "type-token" relations in speech.

As a function having its own duplicating mechanism, speech would not seem to require what is necessary for the functions of our senses, to achieve twin sensations, viz. bi-laterality of the brain function. This is actually the case: speech is a uni-lateral brain function.

17.3.3 The doubling mechanism of language compensates for the uni-laterality of the speech function

Although, or rather because, speech is a uni-lateral function of the brain, linguistic statements occur virtually in duplo. The dual statements are, however, not just repetitions of one another, in which case they would represent one statement only, but show the event—the content—which they are meant to express from different angles.

As far as the dual can be obtained automatically by just substituting "type" for "token" or "token" for "type", the result may—but need not—be trivial from the structural point of view. A typical instance is the formal relation between sentences by virtue of which one sentence structure may be called a transformation of another sentence structure (e.g. the active and passive, or in a different way, question and answer). To show that relation in a completely general way, we use the following notation:

N = noun (N_1, N_2 first and second noun)
V = verb
v = tense (includes -ed and auxiliary verb)

We now state the "active-passive" transformation in English of sentences such as "Children drink milk", and "Milk is drunk by children", in symbols as (Harris, 1957)

$$N_1 \; V \; v \; N_2 \text{ and } N_2 \; v \; V \text{ by } N_1$$

Trivial though the transformation may seem from the point of view of construction, it is by no means trivial from the viewpoint of "content". It shows that neither the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

active nor the passive mood is sufficient for describing the event fully, and that both are needed for that purpose.

Another duality of the automatic variety follows the formula, "rarely true in mathematics, frequently but not always true in logic, that if *a* bears a certain relation to *b*, *b* will bear the same relation to *a*." For instance, to say that a living creature leaves life is using the dual of: life leaves the creature; that the harbor approaches the boat, or that the land recedes from it, are the duals of the boat nearing or leaving the harbor. The description of the corresponding statements as duals will make us beware of the mistake (E. A. Hahn, 1957) of calling the second version "correct" and the first a "fallacy". According to the principle of relativity of motion as formulated by Einstein, there is nothing to choose, as far as "truth" is concerned, between the statements, and that is precisely what duality implies. (For linguistic duality of the non-automatic type, see Chapt. XIX A, Sect. 19.4, and Chapt. XIX B.)

17.4 LINGUISTIC DUALITY AS A METHODOLOGICAL TOOL OF RESEARCH

The importance of the principle of linguistic duality from the theoretical point of view is obvious: among other things, it reveals the relation between the parameters proper to language structure, such as the Characteristic *K* and the Entropy *H*. Its usefulness on the practical side, as a methodological tool of research in structural linguistics, is not so obvious. Below are listed some of its applications to linguistic research.

1. It is useful for resolving controversies, by revealing controversial views as duals, e.g. the views held by Good (1953) and the author (1955) on the stability of the Characteristic *K*. (See above the confronting of the two views in twin columns.)

2. In linguistic research if a relation between type and token is established we may try the reciprocation of these two concepts and see whether the new statement makes sense. If it does, it means enlargement of our knowledge; if not, this may point to the original statement requiring alteration in such a way that its dual is also true.

3. It may provide information about language space, literally and physically. To mention one important result (Vol. I, Sect. 12.4): self-duality of frequency of occurrence of words and frequency of cognition of concepts appears to be the link which holds symbol and thought together, a problem for which, so far, no solution had been offered, but which is vital for the understanding of certain types of speech pathology, e.g. Aphasia (Herdan, 1958).

I trust that the Type-token duality will be considered an improvement in conceptualisation of linguistic duality as first described in Vol. I and the following sections will give a more adequate isolation of the various ways in which duality emerges in language. They will also provide an answer to the question: "What is the good of viewing language in terms of duality?" The answer is very much the same as could be given if the question referred to the advantage of introducing duality in projective geometry, viz. that, residing in the texture of linguistic space, dualities will reveal the nature of that space, and thus contribute to our knowledge of language structure (see Epilogue).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

17.5 THE DISTRIBUTIONALIST PRAGUE SCHOOL CONTROVERSY IN THE LIGHT OF LINGUISTIC DUALITY

In 1884 the mathematician E. A. Abbott had the idea of presenting the relations between different geometries in form of a popular work of mathematical fiction, called *Flatland, a Romance of Many Dimensions.* Since the story is told by "A Square", and thus by a Flatland creature, it gives a true picture of the relations in a two-dimensional world. After a brief sketch of such a world as it appears in the eyes of an inhabitant of Flatland, we shall show that "immanent linguistics" aims at seeing language in a completely analogous way, if the methods of the distributionalist school are strictly followed. This leads to the realisation that the controversy between the distributionalists and the Prague phonological school is one between different geometries, standing in a well-defined relation to one another.

(1) The author of *Flatland* asks us to visualise a world like a vast sheet of paper on which straight lines, triangles, squares and other figures, instead of remaining fixed in their places, move freely about, on or in the surface, but without the power of rising above or sinking below it, very much like shadows. In such a world it is impossible to see what we call "shapes" of figures, since to do so one must rise above, or sink below the surface which, by supposition, is impossible for an inhabitant of Flatland. Nothing is visible for such a one but the edges of the figures as straight lines. The relations between the inhabitants of Flatland, i.e. the population of geometrical shapes, consist thus for them in the mutual positions of lines in a plane. Considering that every shape appears as a distance or as two points on a line, we can say that all the relations between the Flatlanders will be reducible to the duality that two lines lie on a point—which for parallel lines in Euclidian geometry is in the infinite—and that two points lie on a line.

(2) Let us now turn to linguistic facts. The most literal and the most primitive application of that duality to language is the cross-word puzzle, where "point" and "line" simply are literally the points (letters) where the linear sequences of letters (words) cross.

In the original cross-word puzzle we are given two lists of words, "down" and "across", of which we are to find synonyms, assisted by two "rules of the game":

1. Two words as linear sequences may cross each other at certain points, i.e. have a letter in common; and
2. A word is defined, apart from its synonym in the "across" or "down" list, as the distance between its initial and final letter, both included, and thus by two points on a line.

We have thus the condition for geometric duality in the strict sense of the term: the synonyms are to be found subject to the condition, imposed by the geometrical arrangement of the puzzle, that two linguistic "lines" cross one another in a "point", and two linguistic "points" lie on a "line". As the conscious application of the duality device, it may be compared with shorthand, as the conscious application of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

principle of efficient coding, and it could fitly be called the shorthand of duality. The appeal to the general public which this form of entertainment had for many years, until it deteriorated into the mental acrobatics version, testifies to the power which the duality device has over our mind.

Considering now that a word consists of a certain number of letters or phonemes, it may be regarded as a frequency of such units and opposed as "token" to the particular unit which the two words have in common as "type". The point-line duality then becomes one of type-token.

The universe of words considered only from the viewpoint of Distribution may be described mathematically as the total of the type-token relations of language structure. Suitable recording of the phonemes which words have in common, together with the place where they occur, and of their order within words, will represent the complete phonemic system of a language, and geometrically this is nothing but the total of type-token relations in the vocabulary, if by "type" we understand phoneme, and by "token" their sequence within words. Meaning or content of words does not enter, since by supposition, *what one word "perceives" of the other cannot be anything except the point-line relation, that is the type-token duality.*

(3) The inhabitants of Flatland, in spite of not being able to move outside the plane, yet are capable by a system of ingenious inferences of recognising lines as belonging to the same shape, that is as being projections of the same shape. This leads to grouping the lines in Flatland according to certain categories. In terms of type-token, this means that the line is now the type and the number of lines of one particular category the token.

Turning again to the facts of language, we find this change duplicated by the relation of words as types and their frequency of occurrence as token. Incidentally, this shows the linguistic duality between different levels of a language: on the phonemic level, the type is the phoneme, and the sequence of phonemes in a *word is the token*; on the morphemic level, the *word is the type*, and its frequency of occurrence the token. This is the statistical distribution of relative frequencies of different linguistic units, and we have thus passed from vocabulary or phonological to text statistics.

But still, the question of content or meaning has not entered, and it is feasible to carry out investigations about phonemic distribution and vocabulary or morphemic distribution without taking content into account.

(4) However, neither does the story of Flatland end there, nor the views on language structure. The story goes on to describe how one inhabitant, "A Square", incited by algebraic analogies between powers of quantities higher than the second, and geometries of more than two dimensions, succeeds in lifting himself, or rather being lifted from "above", above the plane, and thus becomes bodily aware of what three dimensions mean. In particular, he discovers that he can now see the "inside" of the two-dimensional shapes.

This corresponds to turning from linguistic duality in the plane to that in space, which takes "content" into account. A line in the plane, that is a linear sequence of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

phonemes or letters, is now seen from the "inside" as having a content, and the occurrence of a number of lines belonging to a category as so many incidences of a content.

Moreover, any line in the plane can also be conceived now as being the trace of another plane, and thus as the intersection of two contents.

The controversial views of the distributionalists and the phonologists are thus seen to represent two geometries, a two-dimensional and a three-dimensional one which, however, are by no means contradictory but stand to one another in a well-defined relationship. Linguistic duality has thus led to a reconciliation of two supposedly contradictory views in structural linguistics.

Incidentally, the story *Flatland* is essentially one of the quarrel which arises through one inhabitant, A Square, having come to believe in a 3-dimensional country, which is taboo to the majority of Flatlanders because it is incomprehensible to them. This very aptly duplicates the controversy between the distributionalists—as linguistic Flatlanders—and the Prague School phonologists as linguistic Spacelanders, who believe in "content" as an essential part of phonemic structure.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XVIII A

THE TYPE-TEXT DUALITY

18.1 THE TYPE-TOKEN-TEXT RELATION AS A CASE OF GEOMETRICAL DUALITY IN THREE DIMENSIONS

The principle of duality likewise exists in space. In space a line is determined by two of its points, but it is also determined by two planes which both contain the line. Likewise a plane is determined by three of its points which do not lie in a line. And a point is also determined by three planes which contain this point. Again, two lines which have a point in common, lie in a plane, and similarly two lines which lie in a plane, have a point in common. Such facts suggest the existence of a duality in space, in which a point is dual to a plane and a line is dual to a line. In other words: *points and planes are dual to one another in space, whereas the line is dual to itself.*

This suggestion is fully borne out by a more detailed examination. In general, we can say that to any figure in space consisting of points, lines and planes, there corresponds a figure consisting of planes, lines and points in which to any point of the first figure there corresponds a plane in the second figure, to any plane in the first figure a point in the second figure, and to any line in the first figure a line in the second figure. This is only a very rough description of duality in space, but it is sufficient for our purpose.

Geometrical duality in three dimensions arises in language through the relation between TYPE, TOKEN and TEXT.

As we know, the two versions of the frequency distribution of vocabulary in a literary text, viz., of frequency of occurrence according to different vocabulary items (Zipf) and of vocabulary items according to frequency of occurrence (Yule), are dual to one another, the duality being that of points and lines in a plane.

On a higher level, i.e. when considering word masses from a number of texts, there are again two versions of vocabulary distribution possible. On the one hand, there are the frequency distributions, one for each partial text plus that of the total word mass. On the other hand, there is the distribution of total vocabulary among the different texts, telling us how many words are peculiar to a given text, and how many it shares with 1,2, ... all of the other texts.

The transition from one version of word distribution to the other, be it random or non-random, is always by way of the word frequencies of the total distribution. The vocabulary item, the Type, corresponds, as before, to the point of geometry, and its

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

repeated occurrences, the Token, to the lines; but in addition we have now the Texts, which clearly correspond to the planes of geometry.

In the transition from the frequency distribution to the vocabulary connection of the texts, the frequency of occurrence plays an important role, but disappears in the final product: the description of the word masses in terms of vocabulary connectivity between the texts no longer contains the frequency element. The duality we encounter here is one between words as types and texts, that is in geometrical terms, between points and planes, and corresponds thus to geometrical duality in three dimensions. In full agreement with the geometrical concept, frequency of occurrence which always corresponds to the line, forms the bridge between the dual forms, but no longer appears in the duality, and is, as we say, self-dual.

The three-dimensional character thus follows from the fact that the token element (frequency of occurrence) though the necessary link between type and text sample —just as the line is the necessary link between point and plane—drops out of the ultimate relation between type and text, because it is self-dual, just like the line in geometrical duality.

This relation proves to be of no less importance than that between type and token, and will be seen to provide a definite, positive answer to the question whether meaning is structured.

The application of geometry to language appears now as a sort of "Egg of Columbus". The material of the written language consists of *types* (the letters or their combinations, words), *tokens* (the linear sequences of types) and *texts* (the arrangement of tokens throughout the book), and thus literally of "points", "lines" and "planes". A letter common to two words is their crossing point, and so is a word common to two longer linear sequences; conversely, two letters may belong to one word, or two words to one linear sequence. Furthermore, two pages may have a line or parts of it (words) in common, and two particular words or lines may belong to one page; three pages have, in general, a number of particular letters (points) in common, and, conversely, a page is characterised by the letters and words on it.

The lines of the page do not, of course, cross bodily, nor do the pages themselves cut one another along a line, but their equivalents in the spoken language most certainly do so in our minds, and all phonemic, grammatical and semantic inference rests upon it. It is, therefore, true to say that the linguistic relations are of the nature of those between points, lines and planes in projective geometry.

This development of Part IV adds to the two geometries used in Part II and III a new one. Part I using all three geometries is somewhat outside the system. The geometry of Part II is essentially that of *point masses* or aggregates without arrangement; that of Part III is one of *linear sequences* of unit elements, which is required for dealing with connected pieces of text; that of Part IV is one of *two dimensions* (type, token) as far as it deals with linguistic elements in the plane, and one of *three dimensions* (type, token, text) when dealing with linguistic elements in space.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

18.2 PARTITIONING OF VOCABULARY—UNITY OF STYLE VS. UNITY OF CONTENT

The investigation of the partitioning of the text vocabulary among samples of equal size, and of the number of such samples containing a particular word, revealed the following pattern (Sect. 3.2). By far the largest groups of words are those which are peculiar to a sample, the one-sample words, and those common to all samples, the all-sample words, the latter being slightly, but systematically, fewer in number; a similar systematic difference emerges between the words peculiar to any two samples and those peculiar to any three samples, the latter being again the less numerous of the two.

These quantitative relations are epitomised by the fact that the ratio of the sum of one-sample words and all-sample words to the total sample vocabulary is sensibly the same for all samples within a series and in all three series examined, and must, therefore, be regarded as a characteristic of vocabulary structure.

Moreover, the numbers of the four word categories and their relations show such close agreement with the corresponding numbers and relations as would result on random partitioning of the total vocabulary in a series among four samples, that the conclusion of a random partitioning mechanism, as the ever-present alternative to the conventional hypothesis of a completely deterministic choice of words by the writer, seems inescapable. This implies the generalisation of the observed pattern of vocabulary distribution to any number of samples.

There is, however, some slight deterioration in the agreement between theory and observation, that is between the actual count and the random partition numbers as we proceed from the four samples of one essay by Macaulay to samples from four of his essays and to samples from four different works by Bunyan. Without going into detail, we may say that, by and large, the partitioning of vocabulary among several samples follows a definite pattern, but more strictly so the more homogeneous the style of the samples.

All this under the supposition of a certain homogeneity of content. Since for widely different subjects, e.g. political and scientific, both the vocabulary and its uses change considerably, it is obvious that a significant change in subject must result in exceptional values of V, within a series. What interests us here is the less obvious change in V due to differences in style or authorship in spite of the subject remaining sensibly the same.

The "acid proof" of the theory put forward here that the pattern arises through a certain unity of style, would be if the pattern were found abandoned in text samples from literary works of *different* writers, that is in a series whose samples came from books by different authors.

18.3 DISTRIBUTION OF VOCABULARY IN THE NEW TESTAMENT

The source from which the following tables were extracted is a "statistical dictionary"

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of New Testament Greek, a work of great precision and comprehension, by R. Morgenthaler (1958) (Sect. 1.2). These tables (74-77) give in col. 3 the numbers of words peculiar to each class, and in col. 4 the *average* numbers of words exclusive to one, two etc. samples, and thus for a group of classes. The texts in each of the four series given below are of unequal length, and the V-ratio is therefore calculated, not from the average numbers of the one-sample words, but from the individual numbers for each text sample.

TABLE 74

The Pastoral Letters

Symbol	Vocabulary	Occurrences
A ... 1st Timothy	541	1588
B ... 2nd Timothy	458	1236
C ... Titus	303	658
Total	901	3482

Lines	Word Class	Number of words peculiar to a class	Average
1	(*ABC*)	110	110
2	(*AB*γ)	94	
3	(*A*β*C*)	56	60.3
4	(α*BC*)	31	
5	(*A*$\beta\gamma$)	281	
6	($\alpha B\varrho$)	223	203.3
7	($\alpha\beta C$)	106	

$V_A = 391/541 = 72.2\%$
$V_B = 333/458 = 72.7\%$
$V_C = 216/303 = 71.3\%$

TABLE 75

The Catholic Letters

Symbol	Vocabulary	Occurrences
A ... Epistle of James	560	1735
B ... I Peter	545	1669
C ... II Peter	401	1103
D ... John (I, II & III)	302	2597
E ... Jude	227	456
Total	1271	7560

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Table 75, continued

Lines	Word Class	Number of words peculiar to a class	Average
1	(*ABCDE*)	52	
2	(*ABCDε*)	27	
3	(*ABCδE*)	8	
4	(*ABγDE*)	4	10.2
5	(*AβCDE*)	5	
6	(*αBCDE*)	7	
7	(*ABCδε*)	16	
8	(*ABγDε*)	22	
9	(*ABγδE*)	9	
10	(*AβCDε*)	11	
11	(*AβCδE*)	3	
12	(*AβγDE*)	3	9.3
13	(*αBCDε*)	13	
14	(*αBCδE*)	9	
15	(*αBγDE*)	1	
16	(*αβCDE*)	6	
17	(*ABγδε*)	59	
18	(*AβCδε*)	23	
19	(*AβγDε*)	26	
20	(*AβγδE*)	15	
21	(*αBCδε*)	23	21.7
22	(*αBγDε*)	20	
23	(*αBγδE*)	12	
24	(*αβCDε*)	12	
25	(*αβCδE*)	21	
26	(*αβγDE*)	6	
27	(*Aβγδε*)	277	
28	(*αBγδε*)	263	
29	(*αβCδε*)	165	171.6
30	(*αβγDε*)	87	
31	(*αβγδE*)	66	

Lines of previous table	Number of samples to which word is peculiar	Mean Class Frequency
1	5	52
2-6	4	10.2
7-16	3	9.3
17-26	2	21.7
27-31	1	171.6

$V_A = 329/560 = 58.8\%$
$V_B = 315/545 = 57.7\%$
$V_C = 217/401 = 54.1\%$
$V_D = 139/302 = 46.2\%$
$V_E = 118/227 = 52.0\%$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 76

The Pauline Letters

	Vocabulary	Occurrences	Words peculiar to part	Words shared with all parts
Epistle to the Romans	1068	7094	281	81
I Corinthians	967	6807	246	81
II Corinthians	762	4448	177	81
Galatians	526	2220	92	81
Ephesians	529	2425	83	81
Philippians	448	1624	75	81
Colossians	431	1577	68	81
I & II Thessalonians	471	2296	58	81
Pastoral	901	3484	335	81

For the nine letters, there are nine groups of word classes according to whether a word occurs in one, two ... nine letters, the number of combinations for these groups being

$$9 + 36 + 84 + 126 + 126 + 84 + 36 + 9 + 1 = 511$$

Because of the prodigious length of the unabridged table, we content ourselves with giving the following shortened version:

No. of lines of unabridged table	Number of epistles to which a word is common	Average class frequency	
		Observed	Expected by random partitioning
1	9	81	174
2-10	8	4.67	8.5
11-46	7	1.64	2.1
47-130	6	0.83	1.1
131-256	5	0.80	1.0
257-382	4	1.32	1.4
383-466	3	2.89	3.1
467-502	2	12.80	13.2
503-511	1	157.22	132.5

$V_{\text{Rom.}} = 362/1068 = 33.9\%$
$V_{\text{I Cor.}} = 327/967 = 32.6\%$
$V_{\text{II Cor.}} = 258/792 = 32.6\%$
$V_{\text{Gal.}} = 173/526 = 32.8\%$
$V_{\text{Eph.}} = 164/529 = 31.0\%$
$V_{\text{Phil.}} = 156/448 = 34.8\%$
$V_{\text{Col.}} = 149/431 = 34.6\%$
$V_{\text{Thess. (I,II)}} = 139/471 = 29.5\%$
$V_{\text{Pst.}} = 416/901 = 46.2\%$
$V_{\text{Exp.}} = 306.5/835.7 = 36.6\%$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 77

New Testament

Following Dr. R. Morgenthaler's procedure, the following nine particular texts are considered as making up the total text of the New Testament.

	Vocabulary	Occurrences	Words peculiar to part	Words shared with all parts
Matthew	1691	18278	151	186
Mark	1345	11229	88	186
Luke	2055	19404	332	186
John	1011	15420	101	186
The Acts	2038	18374	562	186
Paul's letters	2648	32303	885	186
Hebrew letters	1038	4942	158	186
Catholic letters	1279	7560	222	186
Revelation	916	9818	127	186

No. of lines of unabridged table	Number of letters to which a word is common	Average class frequency	
		Observed	Expected by random partitioning
1	9	186	584
2-10	8	15.33	22.5
11-46	7	4.56	6.0
47-130	6	2.50	2.9
131-256	5	1.88	2.3
257-382	4	2.92	3.0
383-466	3	6.28	6.4
467-502	2	27.18	25.5
503-511	1	291.75	226

$V_{Mt.} = 337/1691 = 19.9\%$
$V_{Mk.} = 274/1345 = 20.4\%$
$V_{Lk.} = 518/2055 = 25.2\%$
$V_{Joh.} = 287/1011 = 28.4\%$
$V_{Act.} = 748/2038 = 36.7\%$
$V_{Pl.} = 1071/2648 = 40.4\%$
$V_{Heb.} = 344/1038 = 33.1\%$
$V_{CL.} = 408/1271 = 32.1\%$
$V_{Rev.} = 313/516 = 34.2\%$
$V_{Exp.} = 810/2033 = 39.8\%$

18.4 DISCUSSION OF RESULTS—THE STRUCTURE OF MEANING

For all four groups of texts, the Pastoral, the Catholic and the Pauline letters, and the nine books of the New Testament, we find the characteristic pattern of vocabulary partition: the classes of one-sample and all-sample words are the most numerous classes, the former being the larger of the two; the average number of words peculiar to a combination of samples decreasing as the number of samples included in the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

combination increases until a minimum in class frequency is reached for words of the middlemost class; from then on the class frequencies rise again, but in such a way that the classes have slightly greater frequencies than the symmetrically corresponding classes on the other side of the distribution.

The characteristic ratio of the sum of one-sample and all-sample words to the sample vocabulary, which we denote by V (for vocabulary characteristic) varies with the number of degrees of freedom of the sample combinations, but remains sensibly the same within a series. This applies to the three series of letters where it results as

over 70% for 3 text samples (Pastoral letters)
over 50% for 5 text samples (Catholic letters)
over 30% for 9 text samples (Pauline letters)

excepting John's letters in the second series (for which V is on the high side) and the Thessalonians and the Pastoral letters in the third series with V deviating on the low and high side respectively. The New Testament series shows variations in V which we shall discuss presently.

The sensible constancy of V within a series—apart from the exceptions noted above—is most remarkable if we consider that the partial texts are in all three letter series of unequal length. This shows the robustness of the parameter which goes far to make it a really useful instrument for linguistic work where more often than not we are not in a position to cut our material according to strict sampling rules.

Only for the first of the four series can single authorship be claimed. Different authorship applies, on the whole, in the case of the New Testament texts, and can be regarded as established for the Catholic letters; for the Pauline letters it is, at least, doubtful whether they have a single author. Here the exceptional values of V_{Pst} and V_{2Tt} are in excellent agreement with the conclusions reached by Greek Bible scholars.[1]

The lack of stability of V for the New Testament series is very much the proverbial exception to the rule which confirms the rule. We find four texts as having a rather low V, less than 30%. These are the four Gospels, Matthew and Mark having the lowest V of the series: about 20%. A reduction of V can be purely statistical, viz. as the consequence of an undue amount of *identity* between the text samples, which, we know, is true for the two gospels. To understand this, let us think of a number of completely identical texts. For such a series, all word classes are zero except that of all-sample words, since every word would occur in all samples. A limited amount of identity tends to reduce the size of classes of words peculiar to few samples, and specially that of one-sample words. But the number of one-sample words is one of the quantities which make up the numerator of V, and, therefore, partial identity must result in a reduction of V.

The small values of V for the other two gospels—though distinctly greater than for the first two—are also due to the fact that certain parts are the same in all four gospels. The gospel of St. John has the largest V among the four, in accordance with its

[1] Grayston and Herdan, 1959

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

exceptionally small vocabulary (forming the denominator of V); Luke's V is the second largest due to Luke having by far the greatest number of Hapax Legomena and words peculiar to him (being part of the numerator of V) among the four. On the other hand, Paul's letters have an exceptionally high V which agrees with their being not a single text but composed of at least nine partial texts whose range of subjects is different from that of the other parts of the Bible. This probably accounts for the great number of words peculiar to the Pauline letters, and since it is these words which form part of the numerator of V, it explains the larger value of V for these letters. The remaining four texts have values of V between 32% and 36% approximately, which is still considerably below the expectation of V (40%, see next Section). Only the Pauline letters have a V near the expected value. This means that these letters do not show in the vocabulary a special affinity to the rest of the New Testament, but only little more than could be accounted for by random partitioning, whereas other parts, e.g. the Synoptics, show an affinity considerably in excess of what might be expected by random partitioning.

For the Pauline letters the theoretical value of V is 36.6%. The majority of the values of V in this series are not far from expectation, though slightly less, which points to a slight connection above what might be expected by random partitioning only. The Pastoral letters, however, with $V = 46.4\%$, are clearly outside the ordinary run of variation of V, which supports the hypothesis of a different authorship for these letters, and so does the 2nd Thessalonian with $V = 40\%$.

On the basis of these findings, the thesis is put forward that the pattern of vocabulary partitioning as described in Sect. 3.2 is a manifestation of what linguists have recently come to denote by "*structure of meaning*". Insofar, our results are an answer to the question "whether meaning can be said to be structured" (Rulon Wells, 1957).

The pattern of the vocabulary distribution over several samples is independent of the particular content of the texts, and thus fulfils the requirement of complete generality which is indispensible for anything that is claimed to describe structure of meaning as such.

This applies in particular to the statistical parameter V whose magnitude depends only upon the degrees of freedom (the number of text samples) of the sample combinations and upon the basic frequency distribution in the total, but not upon the particular words used, and, therefore, not upon the content. We may express this very important result in a way slightly different from that used so far. Writing the formula (14) p. 62 as

$$V = (A\beta\gamma\delta)/(A) + (ABCD)/(A) \qquad (1)$$

V appears as the sum of two probabilities which, by the addition theorem of probability, means that it represents an alternative probability, namely that of a word being either a one- or an all-sample word. We can, therefore, say: *in texts which may be regarded as "variations upon a theme" the alternative probability of a word being peculiar to one sample or common to all samples is a constant.*

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

It follows that V will vary in a series of texts as are either no longer "*variations* upon a theme", but are, to a greater or lesser extent, identical (see the values of V for the Synoptics: Matthew, Mark, Luke), or which differ too much in style from the rest to be described as "*variations* upon a theme" (see the Pauline Letters as part of the New Testament) and for texts whose content is so different that they cannot be any longer regarded as variations upon *a theme.*

As observed above, the quantity V depends for its magnitude only upon the degrees of freedom of the sample combinations and upon the frequency of word occurrence in the total (though to a lesser extent), but is independent of the particular words used, and with it of the content, thus conforming with the requirements for "structure of meaning" as defined by Rulon Wells (1957).

Although constancy of V *within* a series presupposes a certain homogeneity of content, it is not influenced by a change in content *between* series. It should be well understood that the constancy of V *within* a series depends upon both style and content being sensibly the same; that *between* series is independent from either style or content and depends largely upon the number of parts into which the total was divided, or the degrees of freedom.

The term "structure" is the correct one to use for the property described. In architecture we mean by structure the design according to which the building materials are put together. And, analogously, we speak of the structure of meaning as the design according to which the "building materials" of language are put together.

18.5 RANDOM PARTITIONING OF VOCABULARY AS A STANDARD OF COMPARISON

On the basis of our assumption that appreciable differences in V between the texts belonging to one series—which means for a given number of degrees of freedom—are due to either a certain amount of identity in expression or a lack of unity of style or content, we must regard the essential stability of V as resulting from the random partitioning of vocabulary as implying unity of content or similarity of style. It follows that we could make the agreement of the observed V with its value obtained by random partitioning for the appropriate number of degrees of freedom, the criterion for the texts in the series complying with the description of being "variations upon a theme". An appreciably excessive V over its theoretically expected value we regard as indicating a difference in content or style, a defect of V below its expected value as indicating an identity of expression.

The random partitioning function was derived in Sect. 3.3. Its general form is

$$S f_x \left(\frac{1}{r}\right)^x$$

$$S f_x \left\{ \left(\frac{2}{r}\right)^x - 2\left(\frac{1}{r}\right)^x \right\} \quad \ldots$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$S f_x \left\{ \left(\frac{3}{r}\right)^x - 3\left(\frac{2}{r}\right)^x + 3\left(\frac{1}{r}\right)^x \right\}$$

$$S f_x \left\{ \left(\frac{4}{r}\right)^x - 4\left(\frac{3}{r}\right)^x + 6\left(\frac{2}{r}\right)^x - 4\left(\frac{1}{r}\right)^x \right\} \quad (2)$$

. . . . . .

where r = number of text samples in a series
x = frequency of occurrence of a word in the total frequency distribution
f_x = number of different words with frequency x
S = sign of summation.

For instance, for a series of 7 partial texts ($r = 7$), the series of formulae is as follows:

$$S f_x \left(\frac{1}{7}\right)^x$$

$$S f_x \left\{ \left(\frac{2}{7}\right)^x - 2\left(\frac{1}{7}\right)^x \right\}$$

$$S f_x \left\{ \left(\frac{3}{7}\right)^x - 3\left(\frac{2}{7}\right)^x + 3\left(\frac{1}{7}\right)^x \right\}$$

$$S f_x \left\{ \left(\frac{4}{7}\right)^x - 4\left(\frac{3}{7}\right)^x + 6\left(\frac{2}{7}\right)^x - 4\left(\frac{1}{7}\right)^x \right\}$$

$$S f_x \left\{ \left(\frac{5}{7}\right)^x - 5\left(\frac{4}{7}\right)^x + 10\left(\frac{3}{7}\right)^x - 10\left(\frac{2}{7}\right)^x + 5\left(\frac{1}{7}\right)^x \right\}$$

$$S f_x \left\{ \left(\frac{6}{7}\right)^x - 6\left(\frac{5}{7}\right)^x + 15\left(\frac{4}{7}\right)^x - 20\left(\frac{3}{7}\right)^x + 15\left(\frac{2}{7}\right)^x - 6\left(\frac{1}{7}\right)^x \right\}$$

$$S f_x \left\{ 1 - 7\left(\frac{6}{7}\right)^x + 21\left(\frac{5}{7}\right)^x - 35\left(\frac{4}{7}\right)^x + 35\left(\frac{3}{7}\right)^x - 21\left(\frac{2}{7}\right)^x + 7\left(\frac{1}{7}\right)^x \right\} \quad (3)$$

To simplify the calculation, which becomes rather cumbersome for high values of x, numerical tables have been prepared from which to read directly the probability of a word belonging to a specified number of samples or text parts. These tables are given in Appendix G and cover a range of r from 1 to 30 and x from 1 to 100.

The theoretical figures for a series of 9 text parts for both the New Testament and the Pauline Letters are entered in the last column of Tables 76 and 77. They represent the expected frequencies of words peculiar to a given combination of the parts. Comparing the observed with the theoretical figures, we find a considerable divergence in the first and last items, that is in the frequencies of single- and all-sample words. Considering the inhomogeneity of the books as regards both style and content, this is just what might be expected. If the series were homogeneous, the number of words peculiar to a part would be much smaller, and that common to all

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

parts greater. The sum of the first and last of these quantities is the numerator of the ratio, p. 251; for obtaining the denominator, the figure for a given combination of parts must be multiplied by the number of such combinations containing a particular part; e.g. the expected number of words peculiar to the combination of two parts is 22.5 which figure must be multiplied by the number of combinations of two parts containing a particular part, say *A*.

The last figure is obtained as the number of combinations

$$^{k-1}C_{r-1} = (r-1)!/(k-1)!\,(r-k)! \qquad (4)$$

For the series of 9 parts, the factors by which to multiply the expected numbers per word class belonging to 1,2,3 ... 9 parts are: 1, 8, 28, 56, 70, 56, 28, 8, 1.

Denoting the probability of a word being peculiar to a given combination of parts [expression in curly brackets in formula (2)] by $C(x, r, k)$, the formula then for the denominator of (1), as the expected vocabulary in one part of the text, is

$$\sum \frac{(r-1)!}{(k-1)!\,(r-k)!} \left\{ \Sigma f_x\, C(x,r,k) \right\} \qquad (5)$$

The reader's attention is drawn to the fact that by using formula (5) are we doing essentially what we did in Sect. 4.4 when using formula (24), viz. to calculate the vocabulary to be expected in a given length of text by a law of chance starting from the frequency distribution in the total. But whereas Sect. 4.4 showed only the final result of applying such a method, here we see the mechanism of distribution in all its detail. The combinatorial method of random partitioning is like a slow-motion picture of the summarising procedure when applying the Poisson law, which, indeed, is the virtue of using a law of chance: it saves us much trouble as regards both thinking and computation. But here, for once, we appreciate the slow-motion picture, since it shows so much better what is really happening. It reveals the curiously scrupulous system of balancing the repeated against the unique use of the words when writing different texts on the same subject, or the different parts of a text.

The case is quite analogous to the situation encountered in Chapt. X. There the global value of Lexis' Q as an index for correlation between the two kinds of elements, i.e. between vowels and consonants, was shown to be the result of a complex system of constraints of the Markoff chain type; here, the global probabilities of word occurrence calculated by a law of chance (the Poisson law), are shown to be the result of a complex system of vocabulary distribution by the combinatorial method of random partitioning.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XVIII B

THE STRUCTURE OF MEANING—THE PROSODIC PROPERTIES OF PROSE DICTION

18.6 THE DYNAMIC VIEW OF VOCABULARY PARTITIONING—THE STRUCTURE OF MEANING

So far we have given the statistical view of vocabulary connectivity. According to it, it is the frequency element (Token) which accounts for the appearance of the vocabulary items (Type) in different parts of the text, or more precisely, which accounts for the pattern of vocabulary connectivity of those parts.

However, this is clearly not the way in which we imagine literary works to arise.

The writer makes a plan for his text, say his book, which means visualising the parts required for presenting the different aspects of the subject. On the one hand, he wishes to preserve the unity of the subject which means some degree of repetition of concepts and their linguistic expression: words. On the other hand, if the vocabulary in each of these parts were more or less the same, this would result not only in a great monotony of style, but in an all-too obvious identity of content in the different parts of the text, which would be unsuitable for dealing with different aspects of the subject.

The way out of the difficulty is to adopt a judicious or balanced partitioning of vocabulary, such as is implied in what we have called "the pattern". According to it, the words peculiar to one part of the text, and those common to all parts are by far the most numerous word classes, thus ensuring a high degree of each, unity and diversity of expression, and with it of thought. Furthermore, the words peculiar to any 2, 3 etc. parts are always slightly more numerous than the corresponding classes of words peculiar to all—1, —2, etc. parts, both groups getting smaller as they approach the medium or middlemost class of words.

The number of words peculiar to a combination of text parts or samples is sensibly the same for the same number of parts included in the combination. Thus any combination of 2 parts has sensibly the same number of exclusive words; and so has any combination of 3 parts, and so on.

This is, of course, not meant to say that the writer is aware of performing a vocabulary partitioning of that kind. It is the more remarkable that it should result without a conscious effort on the writer's part.

If what we call the "total", as distinct from the "parts", is not the work of one author, but a composite work of literature, parts of which were written by different authors at different times, each dealing with a different aspect of the subject, a similar balancing mechanism for the distribution of thought and words would have to be at

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

work if repetition and monotony, on one hand, and loss of unity, on the other, are to be avoided. Consequently, we find the same basic pattern emerging in the series of the Catholic and Pauline letters, and, by and large, in the whole of the New Testament, the exceptions being such as to lend support to the theory.

Thus, the primary element are the concepts (Type) which are required for developing a given subject, and it is their use for expressing the relations in which they stand to one another which accounts for their frequency of occurrence (Token) in the text. We have here again the characteristic duality relation between Type and Token. In twin form:

Statistical View	*Dynamic View*
The frequency (Token) as the primary element determines the appearance of words (Type) in different parts of a Text, and with it the pattern of vocabulary connectivity.	The concepts (Type) required for developing a given subject are the primary element whose relations in the manifold of experience determine the occurrence of the corresponding linguistic symbols (Token) in different parts of a Text, and with it the pattern of vocabulary connectivity.

It follows that what from the statistical angle is the random partitioning of vocabulary, is from the dynamic angle the result of a judicious use of concepts according to their relations in nature, or, adopting the term used by linguists, the *structure of meaning*.

In the terminology of Rulon Wells (1958),[1] the duality is that between a complex of events in the plane of expression (the *E*-plane) and that of content (*C*-plane). In the *E*-plane, events occur in linear sequence, and thus in temporal order; "a text is a succession of morphs, and a morph is a succession of phonemes; again, a phoneme is a bundle of simultaneous distinctive features. Temporal order lies at the basis of all these features". In the *C*-plane, on the other hand, there is no exact counterpart to such temporal order. Content or structure of meaning is timeless, since none of the concepts taking part in the pattern could claim to be prior to the others. It is their pattern, or what we call the structure of meaning, which determines the frequency distribution of vocabulary and with it the order in time in the *E*-plane.

Partitioning of vocabulary as a dynamic process designed to show the subject under discussion from all sides, and to achieve the necessary balance between unity and diversity of thought, or—to put it differently—to preserve the unity of the subject while allowing for sufficient diversity in the presentation, requires that the writer should have the TYPE-TEXT relation in all its complexity continuously before his eyes. This is most likely what we call concentrating on a particular aspect of the subject matter, while having the total picture of it at the back of one's mind. This means that

[1] R. Wells, *Proc. 8th Congress of Linguists, Oslo 1958*, p. 659.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the partitioning of vocabulary as a dynamic process requires the constant use of linguistic duality in three dimensions, and is, in fact, its outcome.

The pattern of vocabulary connectivity epitomised by V has the property which is indispensible for what we mean by "structure of meaning": it is perfectly general, in the sense that it is independent of the particular subject of a literary text, and of the style of the particular writer. It is, therefore, a characteristic of meaning in general. It should be observed that it differs in that respect from the Repeat Rate, which is dependent upon both, style and content.

It follows from the nature of random partitioning that the pattern which we have recognised to be the structure of meaning is a pattern of thought, or of concepts, such as to require a certain modicum of vocabulary connectivity, namely just as much of it as results from random partitioning of the total vocabulary. This makes it possible to differentiate between the various kinds of structured meaning, according to whether the observed vocabulary connectivity is equal to, or more, or less than that corresponding to random partitioning. Again, the determination of vocabulary connectivity in excess, or defect, of that due to random partitioning is perfectly general, since it does not refer to a particular subject.

18.7 THE TYPE-CONTENT RELATION AS A CASE OF GEOMETRICAL DUALITY IN FOUR DIMENSIONS

So far we have encountered in language structure geometrical duality in two and three dimensions: as Type-Token relation between the linguistic forms (e.g. words) and their linear arrangement, corresponding to point and line respectively of projective geometry in two dimensions, and as Type-Text relation between the linguistic forms (words) and the occurrence in different parts (number of pages) of a text, corresponding to point and plane of projective geometry in three dimensions.

However, as one proceeds to higher dimensions—mathematically not physically—duality persists, and it enlarges. The dual of a point is not always the same thing, as we have seen, for it is a line in two and a plane in three dimensions. Generally, for a space of n dimensions the last but one space, which is of $(n - 1)$ dimensions and is called a PRIME, is the dual of a point in relation to those n dimensions, so that a prime is a line, plane or solid according as n takes the values 2, 3 or 4 respectively. In this way a chain of dualities arises as we proceed to spaces of higher dimensions. All the smaller dualities belonging to the embedded spaces of fewer dimensions are partial, compared with the fuller duality of the all-embracing space.

We now ask—rather like the hero of *Flatland*—which aspect of language corresponds to the prime in 4-dimensional space, that is, to the volume or solid of geometry.

The answer is not difficult as long as we do not seek it too far: it is the content of the text as reconstructed from the words in our imagination.

Since what we imagine is part of the world around us—even though it may be only imaginary as such—it must be 3-dimensional, and thus agrees in dimensionality with

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the solid of geometry. The reconstruction of what was in the writer's mind thus implies a prime of 3, and with it an embracing space of 4 dimensions. It is therefore true to say that content is the dual of words in four-dimensional space. That this definition of content does full justice to the fact of a lack of one-to-one correspondence between words and content can be illustrated as follows.

A curve, say an ellipse, may be drawn on paper pointwise; but it may also be cut with scissors, which implies its definition by tangents to the curve. Since the tangents as lines are dual or prime to the points we call the analyses the *punctual* and *primal*, respectively. Similarly, in fashioning a work of art in higher-dimensional space, painting is punctual and sculpture primal. The first is the direct, pointwise creation of the work by the brush, the second an indirect fashioning from "without", so to speak.

Quite analogously, writing a book is decidedly punctual, the arrangement of letters, words, etc. being pointwise, one after the other, and each particular one in the right place. Reconstituting the meaning from the text cannot proceed directly since the single disconnected linguistic forms do not tell us what was in the writer's mind. For that purpose we must hold together what the writer has said, not only in connected pieces of text, but in pieces which are a number of words, or even a number of pages apart, and thus use, not single words, but combinations of them, and often in masses. The result is the 3-dimensional creation of our imagination corresponding to 3-dimensional sculpture.

The following paragraphs will describe in greater detail the process of delimitation of content through word combinations.

18.8 FREQUENCY OF OCCURRENCE AS PERIODICITY IN THE USE OF WORDS

In the last section we spoke of the content of linguistic forms as different from the forms themselves, but no definition of content was given. However, the interpretation of the stability of V leads to such a definition which does full justice to the fact that there is no one-to-one correspondence between individual words and content. Content is more, and less, than just the counterpart of discrete words, since it obviously emerges through the interplay of a number of words.

The content of a text necessitates the use of a certain vocabulary. More frequent words will appear more often together in parts of the text than others, and the result of taking the frequencies of all words duly into account when considering their appearance in the text, is a sort of pattern of joint vocabulary occurrence in parts of the text, just as we have seen it actually emerge empirically, and theoretically by random partitioning. We might, therefore, provisionally say that content is what connects different parts. The parts may be any size, that is comprise any number of words, but should for a strict application of the definition be of equal size.

As was shown in the last section, the vocabulary connection of parts of a text is the result of the group of word frequencies, that is of the distribution of vocabulary according to frequency. Considering that frequency of occurrence of a particular

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

word is related to the average gap length between successive occurrences of that word by the formula

$$n/f = \bar{v}$$

where $\bar{v}$ = average gap length
f = number of occurrences of a particular word
n = total text length in terms of word number

it may be described as a periodic function, the word occurring periodically throughout the text with intervals $\bar{v}$.[1] That interval, i.e. the period, is smaller for grammar words, and larger for vocabulary items. Using this interpretation of frequency of occurrence, we may say that *content as the vocabulary connection of parts of a text is the result of, or equals, the group periodicity of words in the total.*

The whole text then appears as a complex system of superimposed periodic or quasi-periodic functions, as far as word occurrence goes. There is, however, something unsatisfactory in this interpretation of content as the result of periodic recurrence of word items. A periodic function has no end, that is, if the iteration of the word is according to a periodic function, then it should go on for ever, whereas content is something permanent, something that is here and now (Eddington).

It was a similar dilemma which led, in physical science, to the introduction of the concept of the *Group Wave* (de Broglie). The problem was as follows: if the velocity of a mass particle (electron) is explained in terms of a system of periodic functions, how can the necessarily finite velocity of the particle be reconciled with the infinite extension of each of the periodic functions? In answer to that question de Broglie has shown that the group velocity of the periodic functions equals that of the particle. In terms of our problem: we ask how the conception of frequency of occurrence of words as periodicity can be reconciled with the permanence of content in a given text; in terms of de Broglie's solution, the answer is that it must be the group frequency which accounts for the permanence of content through vocabulary partitioning, and that is precisely what we have shown to be the case in the last sections.

The Prosodic Properties of Prose Diction

The regularities in the partitioning of vocabulary are capable of an interpretation that will bring the whole matter within the grasp of even the non-mathematically minded linguist. It can namely be shown that the facts of vocabulary partitioning stand to the conventional view about the selection of words in the same relation as Trubetzkoy's phonology to phonetics.

By and large, we can think of frequency of occurrence as implying the periodic recurrence of particular words. The whole machinery of diction in the sense of selection of words could then be conceived as a complex system of periodicities, one for each word, but comprising so many identical waves as there are vocabulary items

[1] With the restriction mentioned below.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

with the same frequency of occurrence. It is the group solidarity of all these functions which brings about (by random partitioning) the permanence of content.

Apart from rhythmical prose, it was so far only in poetic diction that periodicity of linguistic forms was known: as the rhythmical iteration of stress or accent of sounds (alliteration) and morphemes (rhyme) and, especially in the classical languages, of quantitative characteristics, long and short syllables.

The above interpretation of random partitioning of vocabulary amounts to an extension of the prosodic principle of iteration, in the sense of periodic recurrence of linguistic forms, to prose in general, with this modification, however, *that the prose periodicity is not one of either sound or quantity, but of whole words.* Content thus becomes the manifestation of word waves, and we can say that *content is the permanence of the group wave composed of the fleeting waves of individual words.*

This is analogous to Trubetzkoy's step from phonetics to phonology. What he did may be described as extending a principle of grammar, namely, that of differentiation of meaning among the items of a paradigmatic series by changing the phonemic components, to the distinction of words in general, that is of vocabulary items, and particularly of word roots which are outside the realm of grammar.

Quite analogously, we now extend the use of a principle of poetry—periodicity of occurrence of certain linguistic elements—to prose of every kind, and to the use of words in general. Trubetzkoy's extension belongs to the two-dimensional aspect of language, or the type-token relation; ours to the three-dimensional, or the type-text relation.

More precisely, we may say that poetry emphasises the content structure of the apparently uniform linear sequence of linguistic forms by periodic recurrence of the same sound (rhyme) or the same quantity (metre) (Sect. 19.11). *The resulting agreement between equidistant parts of the poetic line corresponds to the agreement between different parts of the prose text line through sharing the same word*, the by and large periodic recurrence of grammar forms—even in prose—giving way to *the incompletely periodic pattern of vocabulary connectivity* through non-grammar words.

As a rule, it is only the poet who is believed to be governed by language more than he governs it. Goethe describes in *Dichtung und Wahrheit* how in his youth he would wake up in the middle of the night, jump out of bed and without sitting down at his desk scribble across a piece of paper a poem that had just come into his head, having had the experience that even a little delay might obliterate it from the excited "tablets of his mind", and he adds that in those days he considered his relation to his poems —"seine Geisteskinder"—to be very similar to that of a hen to her chicks.

It is now our contention that the diction of good prose is not an affair of hand-picking of words by the writer, but that, to a large extent, the writer is, so to speak, at the mercy of language, just like the poet. He is, of course, master of the content of the whole and its parts. But once the system of content has been settled in outline not only the main items of the vocabulary needed for expression, but also their relative frequencies are, by and large, determined, the system of content determining how often a particular word will appear in combination with certain other words.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

This means that in the execution of his task in detail the writer is no longer free, since words will have to appear more or less periodically, each with its own period, and thus be "at the right time in the right place" for unexpected combinations with other words. Words will then act as "prompters" and suggest ideas to the writer. Thus, although in the *strategy* of writing the writer is free, he is no longer so in the *tactics* where he cannot escape the consequences of his own plan of action: or briefly, whereas the strategy of the "Battle of Words" is the work of the writer, its tactics are the work of language. Nobody has expressed this better than Karl Kraus.[2]

This may seem to accord to chance (as involuntary action) in speech an inferior role, and reminds one of the *de minimis non curat lex*, *lex* being the system of content determined by *choice*, whereas the *minima* are the features of the text left to *chance*. However, Butler rightly asks: "Yes, but what is a minimum? Sometimes a maximum is a minimum, and sometimes the other way about. If you know you know, and if you don't you don't." And he adds: ". . . it seems as though the maxim '*de minimis non curat lex*'—the fact that a wrong, a contradiction in terms, a violation of all our ordinary canons does not matter and should be brushed aside—it seems as though this maxim went very low down in the scale of nature, as though it were the one principle rendering combination (integration) and, I suppose, dissolution (disintegration) also, possible. For combination of any kind involves contradiction in terms; it involves a self-stultification on the part of one or more things, more or less complete in both of them . . . It is this principle which is embodied in *de minimis non curat lex*, that makes it possible that there should be *majora* and a *lex* to care about."[3]

With this interpretation, the Latin legal maxim appears to be the forerunner of the principle of experimental design for purposes of variance analysis. The *common error term* which we extract there from the experimental data is not all "error", but comprises the variation due to all causes or factors in which we do not happen to be interested—for the moment, that is—and which have, therefore, not been included as possible factors in the design of the experiment. The common error term is thus only a *relative minimum*, but as such determines the magnitude of the variation due to the *majora* (the factors); moreover, it contains very important information about the possible interaction of the factors, and a good deal of the analysis consists in scaling down the error term to its minimal size by extracting from it as much information as possible about the factors.

[2] *Die Sprache*, 2nd ed. (Munich, 1954).
[3] Samuel Butler, *The Note-Books* (London, 1930).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIX

MATHEMATICAL PHILOLOGY

19.1 LANGUAGE ABOUT LANGUAGE

So far the application of mathematics has been to linguistics, apart from stylostatistics which might be said to belong to the realm of philology. However, even there it was only formal characteristics that were subjected to mathematical analysis and no attempt was made to use quantitative methods for assessing literary quality, a matter which sooner or later must enter philological work.

It stands to reason that if such methods are useful for describing and analysing linguistic events which are the raw material of works of literature, they should also prove helpful in describing and analysing the latter. This, as we shall see, is actually the case, and here we enter upon a new field of application of mathematics to language. Be it said at the start, the methods which are useful here are not those of the mathematics of measurement (statistics), but those of the mathematics of order and position (projective geometry). The new application will come as a welcome relief to the linguist who has worked his way so far and may sympathise with Darwin when he laments in a celebrated passage the atrophy of his aesthetic appreciation and confesses that his mind "seems to have become a kind of machine for grinding general laws out of large collections of fact".

The recent development of linguistic theory has been aptly described as linguistics hauling itself up by its own shoe strings. It has to some extent become language about language about language and so on indefinitely. Insofar as this is really practised, the result is not infrequent that the investigator ties himself up into quite complicated knots. The type of problem which the modern linguist tackles clearly requires other and more powerful concepts than those at his disposal. The contention put forward here is that such concepts are provided by a certain branch of mathematics, viz., that of projective geometry. The most powerful conceptual tool in mathematical philology is the same which has proved most useful in mathematical linguistics: duality. There is, however, this difference: there it was the substitution of TYPE and TOKEN for point and line respectively which transformed the aprioristic geometrical relation into one between observed linguistic events; here, where linguistic structure is replaced by text structure, duality in space takes the place of duality in the plane, and we shall have to find the style characteristics which, when substituted for "point" and "plane" will transform geometrical duality into one of philological events.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIX A

THE CRAFT OF SATIRE

19.2 THE DECAY OF LANGUAGE THROUGH THE ABUSE OF LINGUISTIC DUALITY—A LINGUISTIC ANALYSIS OF NAZI GERMAN

The language which developed in Germany after the 31st of January 1933 had two outstanding features: "Gleichschaltung" and Duplication. The first was the vain attempt to identify linguistic symbol and reality, signifiant and signifié, expression and content, once and for all. The latent German tendency to be satisfied with words and not trouble too much about the reality behind them (Goethe: "... denn eben wo Begriffe fehlen, da stellt ein Wort zur rechten Zeit sich ein", *transl.* "... for it is just when concepts are missing that words come pat") was here pushed to its utmost extreme. The word was hypostasised into a sort of self-evident unit which no longer derived its currency value, its raison d'être, from the part of the real world it was meant to denote, but which from now on was to be used with such meaning as suited the aims of the NSDAP, the SA and SS. Certain words, like Jud, Blut, Gut, Rasse, Deutsch, were chosen as key words and identified with everything good and noble, like Deutsch, or evil and bad, like Jud, which resulted in their endless repetition to the virtual exclusion of the greater part of "neutral" vocabulary. From now on it was the arbitrary, and mostly criminal, intention of the speaker that decided in which relation the words were to stand to one another, and no longer their solidarity as parts of "la langue".

The need for the second feature, duplication, arose in this way. To dictate how a word was to be used, and with what content, proved acceptable inside Germany, where opposition was unthinkable because punishable by "brevi-manu"-extinction, but failed to work vis-à-vis the outer world whose wayward inhabitants could not be induced to accept every pronouncement by the "Führer" and his underlings as Gospel truth, especially whenever it was in complete contradiction with the facts—which it happened to be more often than not.

To overcome this difficulty, the German mind had resort to an expedient which, in a way, was contrary to "Gleichschaltung", although the Nazi mentality succeeded in modifying it until it fitted into their idea of what a language should be like, or what it was meant for. And this is for us the most remarkable feature of Nazi German: by showing the profound truth expressed by Goethe that we obey the laws of nature even when we try to act against them (*Essay on Nature*, 1788), it establishes better than

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

anything else the fact of duplication as a consequence of linguistic duality. In Nazi German this took the following form: "*Gleichschaltung*" *as a straight-jacket of language made it necessary to adopt for relaxation purposes a perversion of linguistic duality.*

This is a matter very little understood so far, and should interest social psychologists and historians as well as linguists, for it explains the Nazi mania of trying to penetrate into the Jewish mind while destroying Jewish bodies. In those days there appeared in Germany a number of newspapers or journals—of the "Stürmer" type—which quite apart from inciting the masses to atrocities against the Jews, were constantly occupied with analysing what they thought was the Jewish mentality, clearly with a view to profiting by it. However conceited the Nazis were, judging by their public utterances, they had little illusion about the fact that the Jews were their superiors in the use of language; consequently, they tried hard to acquire such versatility in the use of that important intellectual tool as seemed to come so natural to the members of the despised race. This, they fondly imagined, would enable them to diddle the outer world.

Doubling of linguistic statements through the type-token duality as one of the most effective devices of the art of language came not only as a revelation, but as a Godsend to the Nazis who saw in it the very thing needed in their discourse with the outer world. A striking instance of "abusus optimi pessimus"! What to the evangelists came as a revelation because it gave them deep insight into the true structure of human relations,[1] proved for the Nazis to be the perfect tool for dissimulation of the truth—at least in the way in which they applied it. Instead of using the doubling device as a means for getting nearer to the truth, they misused it for the purpose of concealment and for counteracting any accusation and condemnation of their behaviour, no matter how true and justified, by simply reversing the role of active and passive part, the resulting logical centaur being what K. Kraus called the "verfolgende Unschuld" (the "persecuting innocence"). This development executed with truly German thoroughness led to the every-increasing orgy of lies in the German utterances from 1933 onwards. But disastrous though this was for the particular language, German, *it provides today invaluable and irrefutable evidence of linguistic duality as one of the inescapable laws governing the correct use of words.*

19.3 THE CRAFT OF SATIRE

The obvious way for substantiating the contention about the misuse of the doubling principle of language by the Nazis would be to quote from the German newspapers of that period. However, thanks to Karl Kraus' *Die Dritte Walpurgisnacht* (1952) we are saved that trouble, and with it the nausea of a perusal of Nazi writings.

This book contains not only all the documentary evidence we require, but in such a

[1] See Sect. 19.4.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

form as to show the doubling principle in operation. Not being a linguist, Kraus had no reason to stress that feature—even had it been known at that time—for its own sake, but he does so as a satirist. One of the main characteristics of the satirist's style is the exaggeration of both, the misdeeds and their excuses put forward by the culprit, until the criminal appears as the innocent hero of the piece and the victim as the villain. Now, in Hitler's Germany, the words and deeds were such as to exaggerate beyond belief even the satirist's inventions, which explains the opening words of Kraus's book: "Mir fällt zu Hitler nichts ein." ("Hitler left nothing for me to say").[2] While the first and second Walpurgis Nights were Goethe's *inventions*, the third consists in matter-of-fact *reality* which, according to Kraus, frustrated human imagination by a total exhaustion and perversion of its resources.

To any incriminating statement about Nazi behaviour the Nazi press itself provided the counterpart, the dual, by way of either justification or denial of the original. All that was left for Kraus to do was to show these statements side by side, almost after the twin fashion of genuine duals, and the satire was there. It was, therefore, natural that the method used by Kraus for describing Nazi mentality should take the form of what I have called linguistic duality. The following is a list of such dual statements from Kraus's book, which may be regarded as an incomparable object lesson of the inescapability from linguistic duality in both the use of language (by Kraus) and its abuse (by the Germans).

"Jene verfolgende Unschuld, die Einheit von Schuld und Lüge, wo die Tat zum Alibi wird und der Greuel zur Glorie."

"Der Circulus vitiosus und perniciosus, worin die falsche Kausalität waltet."

"Ein Lied (composed by the Nazis) mit dem Title und Refrain, 'Juden raus', eine Forderung, die der Tendenz gegen Kapitalsflucht freilich durchaus entgegensteht."

"Der Nationalcharakter, dessen Vertreter 'Juden raus' rufen und zugleich dem Versuch der Befolgung an der Landesgrenze entgegentreten."

"In der Fähigkeit Blamage in Effekt umzuwechseln, kurz darin was man früher Blödmachen nannte, ist das nationalsozialistische Kommuniqué vorbildlich."

"... *dass man das eigene Wort als üble Nachrede empfindet, wenn's ein anderer glaubt.*"

"Der Nationalsozialismus hat überhaupt keine andere Waffe als den umgekehrten Spiess."

"... *immer die gleiche Antinomie der Erscheinungen...*"

"Man weiss von nichts und redet von etwas anderm, man hat nichts getan, aber der andere ist schuld; es ist nichts geschehen und er hat es getan; *man bezichtigt den, der die Wahrheit sagt der Lüge, auf der man ertappt wurde.*"

"... eine Denkart, die mit totaler Naivität Handlungen setzt, die man sonst erst als Karikaturen erfinden müsste."

"Eine gewisse Erleichterung des Verkehrs wäre zunächst schon angebahnt, indem

[2] Strangely reminiscent of Faust's words: „Der Teufel hat hier weiter nichts zu sagen" (*Faust II*, 2nd act).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

man einfach vorweg annimmt, dass das Gegenteil gemeint sei, wiewohl auch das nicht sicher ist."

"Und wahraftig, soweit die deutsche Doppelzunge klingt, empfängt die hörende Welt mehr Überzeugung, als sie braucht."

Translation into English

"That 'persecuting innocence' (of the Nazis), the unity of guilt and lie, where the act is used as an alibi and the atrocity is glorified."

"The vicious and pernicious circle of the false causality."

"A song (composed by the Nazis) with the title and refrain 'Out with the Jews', the compliance with which request, however, exposes the victims to the charge of 'flight from the Reichsmark'."

"The national character whose representatives shout 'Out with the Jews', and at the same time obstruct on the border any attempt to comply with that request."

"In the tendency of turning blamage into effect, and briefly in what is commonly called 'making fools of people', the Nazi communiqué cannot be surpassed."

"*They feel their own word as slander, if somebody else believes it.*"

"National-socialism has no other weapon than the spear turned round."

"... *always the same antinomy of events* ..."

"One knows of nothing, and speaks of something else; one hasn't done anything, but it was the fault of the other man; nothing has happened, and he has done it; *one imputes to him, who tells the truth, the lie which one was found out to have used oneself.*"

"A way of thought which with complete naivety pretends situations which usually are classed as caricatures."

"A certain simplification in one's relation with the Nazis would be to assume systematically that they always mean the opposite of what they say, though even that cannot be depended on."

"And truly, as far as the German Double-Tongue reaches, the listening world receives more information than it would need."

19.4 THE POWER OF LINGUISTIC DUALITY

By way of contrast with the travesty of linguistic duality perpetrated by the Nazis, there are linguistic dualities which are anything but automatic and arbitrary, but which, just like certain geometrical dualities, afford the most unexpected insight into recondite relationships. They are true revelations of content. A good illustration is provided by certain passages—and they are the most important ones—from the Sermon on the Mount. Christ's peculiar style is apparent in the opening sentences: using straight and wavy underlining to denote the duality between such statements, the duality is seen to be between true meaning (*literalism*) and extended meaning.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

"Blessed are the poor in spirit: for their's is the kingdom of heaven.
"Blessed are they that mourn: for they shall be comforted.
"Blessed are they which hunger and thirst after righteousness: for they shall be filled.
"Blessed are they that are persecuted for righteousness sake: for their's is the kingdom of heaven."

Each of these sentences is the dual of the corresponding conventional view:

"*Blessed are the rich in material goods*: for their's is the kingdom on earth.

"*Blessed are they that have no reason for mourning*: for they require no comfort.

"*Blessed are they who don't suffer hunger*: for they are satisfied.

"*Blessed are they who are free and don't suffer persecution*: for their's is the earth."

In the text, Christ continues to use this device, and it is of great significance, and testifies to its power, that the most striking passages are put in that form, e.g.

"... Resist not evil: but whosoever shall smite thee on they right cheek, turn to him the other also.
"... Love your enemies, bless them that curse you, do good to them...
"... Therefore take no thought for the morrow: for the morrow shall take thought for the things of itself. Sufficient unto the day is the evil thereof."

These commandments are dual to those which the Jews in their struggle for existence with the neighbouring warlike nations had naturally to follow.

The realisation of the power of language to express the most recondite relations between the world and ourselves through the use of the doubling device afforded by linguistic duality must have come like a revelation to those who stored in their minds, and later recorded in writing, the word of Christ. This may be the explanation—and a truly immanent one—of the opening sentences of St. John's Gospel:

"In the beginning was the Word, and the Word was with God, and the Word was God."

The third sentence in the Authorised Version attributes creation to God, but Luther's translation attributes it to the Word:

"Alle Dinge sind durch *das*selbe gemacht, und ohne *das*selbe ist nichts gemacht was gemacht ist."

Luther seems to render more correctly the idea of the writer, since he makes the third sentence continue the first sentence, whereas the Authorised Version by attributing creation to God, not to the Word, leaves the opening sentence hanging in the air.

By putting God = Word = Life = Light, the Evangelist contrives to accord to the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Word the role of that which transforms darkness into light, and with it chaos into order. The Word thus appears as the ordering element. As we today know, this is quite correct from the viewpoint of structural linguistics, and the most general order device of the Word is duality resulting in the duplication of the number of linguistic statements. Since it is through the element of order peculiar to the Word that the World is apprehended, understood and overcome, it is perfectly proper to say that "the Word was in the beginning ... and the Word was God".

As another milestone of literature consisting almost entirely of dualities may be quoted *Goethe's* celebrated essay *Die Natur*. The following are disconnected passages:

"Natur! Wir sind von ihr umgeben und umschlungen—unvermögend aus ihr herauszutreten, und unvermögend tiefer in sie einzudringen.

"Sie schafft ewig neue Gestalten: was da ist, war noch nie; was war kommt nicht wieder; alles ist neu und doch immer das alte."

"Sie scheint alles auf Individualität angelegt zu haben, und macht sich nichts aus den Individuen; Sie baut immer und zerstört immer, und ihre Werkstätte ist unzugänglich."

"Sie lebt in lauter Kindern, und die Mutter wo ist sie?"

"Jedes ihrer Werke hat sein eigenes Wesen, jede ihrer Erscheinungen den isoliertesten Begriff, und doch macht alles eins aus. Es ist ein ewiges Leben, Werden und Bewegen in ihr, und doch rückt sie nicht weiter."

"Man gehorcht ihren Gesetzen, auch wenn man ihnen widerstrebt."

"Sie hat keine Sprache noch Rede, aber sie schafft Zungen und Herzen durch die sie fühlt und spricht."

"Sie hat alles isoliert, um alles zusammenzuziehen."

"Sie ist alles. Sie belohnt sich selbst, und bestraft sich selbst, erfreut und quält sich selbst."

"Sie ist ganz, und doch immer unvollendet... Jedem erscheint sie in einer eigenen Gestalt. Sie verbirgt sich in tausend Namen und Termen und ist immer dieselbe."

"Alles ist ihre Schuld, alles ist ihr Verdienst."

Translation into English[3]

"Nature! We are entwined and surrounded by her, incapable of escaping from her, incapable of penetrating more deeply into her."

"She creates continuously new forms; what is, was not before; what was, does not come again; everything is new, and yet still the old."

"She seems to aim only at individuality, and yet does not care about individuals. She constantly builds and destroys, and her workshop is inaccessible."

"She seems to live only in her children, and the mother, where is she?"

"Everyone of her creations has its own individuality, each manifestation the most

[3] In this translation, the pronoun "She" is used for "Nature" in order to personify her, as in Goethe's essay.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

isolated content, and yet all is one. There is eternal life, development and movement in her, and yet she does not progress."

"We obey her laws, even when we oppose them. We work with her, even when we work against her."

"She has no language, no speech; but she creates tongues and hearts through which she feels and speaks."

"She has isolated all things, in order to unite them."

"She is everything. She rewards herself, and punishes herself, pleases and tortures herself."

"She is whole, and yet always unfinished... To everyone she appears in a different shape. She hides herself under a thousand names and terms, and is always the same."

"All is her fault, all is her merit."

That the antithetical style is not only due to the poet's description of nature, but to the complexity of the manifold, can be inferred from the fact that modern theoretical physics with all its apparatus of mathematical symbolism is one vast system of dualities: as atomic physics, vs. wave theory, deterministic vs. statistical physics, empirical vs. aprioristic physics (Eddington, *Fundamental Theory*, 1947).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIX B

THE CRAFT OF FICTION

19.5 DRAMA VS. FICTION

The influence of the linguistic (conceptual) and mathematical (symbolic) codes upon the formation of physical theories, and with it upon our picture of the physical world, has been the object of many investigations in recent years. The equally interesting problem of the influence of the linguistic coding devices upon our picture of the world of fiction, that is virtually of the life of our species, has received only scant attention, the most penetrating study, Lubbock's *Craft of Fiction*, dating back to 1921. Our literary critics have been oddly incurious in the matter. This is proved by the singular poverty of our critical vocabulary, as soon as we pass beyond the simplest effects. The expressions and phrases at our disposal for the description of fiction, such as "dramatic", "narrative", "pictorial", "scenic", bear no definite meanings. We must use such words as soon as we begin to examine the structure of the novel, and yet they have no technical acceptance with regard to the novel. The want of an established nomenclature is a real hindrance in any effort to understand and describe the structure of the novel.

As will be seen, the theory of language developed in this book provides the apparatus for describing novel structure, and for understanding the craft of fiction, in a more satisfactory way than has been achieved hitherto.

In speaking of the "form" of the novel, we use the term in a metaphorical sense. The story or the "information" is derived from a linear or one-dimensional sequence of words, which represents the physical form of the novel. Our problem is to explain how the sequence of letter groups, called words, is capable of communicating to us the whole complexity of human life. The way of arranging the material so as to convey the multi-dimensional complexity of life by means of the one-dimensional channel of words, this is what we call the form of the novel.

The novelist, as different from the dramatist, uses exclusively a linguistic code. In the drama, the action is developed by means of things that remain outside the coding device; by means of real things, that is, and not by conventional symbols of things. Real live men and women move about the stage; we hear real voices. The things remain as they were in real life, and are not transmuted into any conventional symbolic form. It follows that for the dramatist there result many limitations in time and space.

When we turn to romance, we no longer find this. There nothing is reproduced to

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

our senses directly. Everything is brought home to us through the medium of another man's mind, and through that of the language symbols, that is, in the form of written words. But with the loss of realism, there is for art a clear gain in liberty and power of expression of subtle relations, just as painting, a two-dimensional picture of the manifold, is far more free and versatile than sculpture which preserves the three-dimensional aspect. *It is by giving up these identities that art gains true strength*, and presentation becomes more "true". Continuous narration is the flat board on to which the novelist throws everything. Even when he uses scenic presentation, his only means of communication remains the word. From this there results a loss of vividness, but also great compensating gain in his power over the subject. He can now subordinate one thing to another according to importance, and introduce all manner of subtle detail, to a degree that was before impossible.

However, the discrepancy between the code symbols and the manifold of real objects has a serious drawback, as hinted above, when we spoke of the loss of vividness. There is a certain thinness in a mere symbolic report of things, for thin and flat must be the reflection we receive from the mind of another, and in terms of uncongenial material like words. It is the novelist's consciousness of this which leads to what may be called the devices of the craft of fiction, as Lubbock has shown. There is a constant effort throughout the course of fiction to counteract the inherent weakness of the method that a story-teller is bound to use, and that indeed is peculiarly his, and which compels him to depend upon nothing but the linear sequence of words.

19.6 THE CRAFT OF FICTION—DUALITY IN FIVE DIMENSIONS

In fiction, too, the content is reconstituted from the words, but it is a peculiar kind of content, namely one which we know to be fictitious. This means that we cannot separate it completely from the story-teller, be it the author or a character inside the story. It is *content plus its provenancy*.

We must, therefore, accord to the reconstituted image of reality another dimension, and, since the image in itself is three-dimensional, regard the fiction content as four-dimensional. It follows that the space of fiction embracing a four-dimensional PRIME (fiction content) as the dual of TYPE (words) has five dimensions. Quite apart from its intrinsic interest for understanding *literary structure* and bringing it into line with *linguistic structure*, our interpretation provides the terminology so sorely missed by Lubbock for adequately describing and analysing the Craft of Fiction. The whole intricate question of method in fiction is governed, according to Lubbock, by that of the point of view—namely the relation in which the narrator stands to the story. He may tell it as *he* sees it; or he may introduce another point of view by setting up a new narrator: the story-teller may be in the story himself, and nothing is now imported into the story from without; it is self-contained.

In the light of language geometry, this shifting of the viewpoint is the alternation between the *punctual* and the *primal* description, the first being the direct presentation

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XIX B

THE CRAFT OF FICTION

19.5 DRAMA VS. FICTION

The influence of the linguistic (conceptual) and mathematical (symbolic) codes upon the formation of physical theories, and with it upon our picture of the physical world, has been the object of many investigations in recent years. The equally interesting problem of the influence of the linguistic coding devices upon our picture of the world of fiction, that is virtually of the life of our species, has received only scant attention, the most penetrating study, Lubbock's *Craft of Fiction*, dating back to 1921. Our literary critics have been oddly incurious in the matter. This is proved by the singular poverty of our critical vocabulary, as soon as we pass beyond the simplest effects. The expressions and phrases at our disposal for the description of fiction, such as "dramatic", "narrative", "pictorial", "scenic", bear no definite meanings. We must use such words as soon as we begin to examine the structure of the novel, and yet they have no technical acceptance with regard to the novel. The want of an established nomenclature is a real hindrance in any effort to understand and describe the structure of the novel.

As will be seen, the theory of language developed in this book provides the apparatus for describing novel structure, and for understanding the craft of fiction, in a more satisfactory way than has been achieved hitherto.

In speaking of the "form" of the novel, we use the term in a metaphorical sense. The story or the "information" is derived from a linear or one-dimensional sequence of words, which represents the physical form of the novel. Our problem is to explain how the sequence of letter groups, called words, is capable of communicating to us the whole complexity of human life. The way of arranging the material so as to convey the multi-dimensional complexity of life by means of the one-dimensional channel of words, this is what we call the form of the novel.

The novelist, as different from the dramatist, uses exclusively a linguistic code. In the drama, the action is developed by means of things that remain outside the coding device; by means of real things, that is, and not by conventional symbols of things. Real live men and women move about the stage; we hear real voices. The things remain as they were in real life, and are not transmuted into any conventional symbolic form. It follows that for the dramatist there result many limitations in time and space.

When we turn to romance, we no longer find this. There nothing is reproduced to

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

our senses directly. Everything is brought home to us through the medium of another man's mind, and through that of the language symbols, that is, in the form of written words. But with the loss of realism, there is for art a clear gain in liberty and power of expression of subtle relations, just as painting, a two-dimensional picture of the manifold, is far more free and versatile than sculpture which preserves the three-dimensional aspect. *It is by giving up these identities that art gains true strength*, and presentation becomes more "true". Continuous narration is the flat board on to which the novelist throws everything. Even when he uses scenic presentation, his only means of communication remains the word. From this there results a loss of vividness, but also great compensating gain in his power over the subject. He can now subordinate one thing to another according to importance, and introduce all manner of subtle detail, to a degree that was before impossible.

However, the discrepancy between the code symbols and the manifold of real objects has a serious drawback, as hinted above, when we spoke of the loss of vividness. There is a certain thinness in a mere symbolic report of things, for thin and flat must be the reflection we receive from the mind of another, and in terms of uncongenial material like words. It is the novelist's consciousness of this which leads to what may be called the devices of the craft of fiction, as Lubbock has shown. There is a constant effort throughout the course of fiction to counteract the inherent weakness of the method that a story-teller is bound to use, and that indeed is peculiarly his, and which compels him to depend upon nothing but the linear sequence of words.

19.6 THE CRAFT OF FICTION—DUALITY IN FIVE DIMENSIONS

In fiction, too, the content is reconstituted from the words, but it is a peculiar kind of content, namely one which we know to be fictitious. This means that we cannot separate it completely from the story-teller, be it the author or a character inside the story. It is *content plus its provenancy*.

We must, therefore, accord to the reconstituted image of reality another dimension, and, since the image in itself is three-dimensional, regard the fiction content as four-dimensional. It follows that the space of fiction embracing a four-dimensional PRIME (fiction content) as the dual of TYPE (words) has five dimensions. Quite apart from its intrinsic interest for understanding *literary structure* and bringing it into line with *linguistic structure*, our interpretation provides the terminology so sorely missed by Lubbock for adequately describing and analysing the Craft of Fiction. The whole intricate question of method in fiction is governed, according to Lubbock, by that of the point of view—namely the relation in which the narrator stands to the story. He may tell it as *he* sees it; or he may introduce another point of view by setting up a new narrator: the story-teller may be in the story himself, and nothing is now imported into the story from without; it is self-contained.

In the light of language geometry, this shifting of the viewpoint is the alternation between the *punctual* and the *primal* description, the first being the direct presentation

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of the content by the writer as the story-teller selecting the words (points) and arranging them (syntactically) one by one, the second being the indirect presentation of the content by letting it arise in the reader's mind from the reports of the story which, since these reports come, by supposition, from people who are not engaged upon novel-writing, are tangential to the image which the author wishes to be reconstituted by the reader from the words.

Expressing the matter in this way has not only the advantage of precision, but it guards us from the mistake which Lubbock made when he ascribed superiority to the indirect method of telling a story.

According to the concept of geometrical duality, both members, point and prime, are of equal rank or importance for the description of the geometrical relation or event in question. Similarly, we must regard the linguistic point and prime in fiction, namely the words chosen by the author as story-teller and the words pretended to be those of a person inside the story, as being both equally necessary, and thus of equal value for the creation of the fiction content. There is thus no reason why the latter should be valued higher. Lubbock's evaluation arose from the mistake that the purpose of fiction was to make the reader forget that he was reading fiction. This is, however, not so: the reader of a novel is always aware that what he is reconstituting from the words is a content or image in somebody's head, and this adds to the appeal of fiction. It is an essential characteristic of reading fiction, as different from reading reports of facts. The geometrical interpretation is in full agreement with this.

19.7 PHILOLOGICAL DUALITY

Mathematical philology has had a forerunner in Percy Lubbock's *Craft of Fiction* He stresses the fact that in all our talk about novels we are hampered by our own unfamiliarity with what is called their technical aspect, and believes that if only there were a single tangible and measurable fact about a book it would be a support in a world of shadows. As such a tangible fact he regards the relation between the viewpoint of the story-teller and the plane upon which the story unfolds itself, especially the novelist's reluctance to *tell* his story and his desire to interpose another presence between himself and the reader. In what follows, the essentially mathematical nature of this change in relationship between point of view and plane of action will become apparent, and this realisation will enable us to improve upon Lubbock's analysis of the craft of fiction.

The whole intricate question of method in the craft of fiction is governed by that of the point of view from which the canvas of life is seen, that is by the relation in which the narrator stands to the story.

He may tell it as *he* sees, it, that is, as one standing outside and above events; which is, no doubt, the most direct way of telling a story.

He may put himself into the story as the story-teller, with or without disguise; nothing is now imported into it from without—or so it appears.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

By a repetition of the same stroke, the viewpoint may be further shifted to characters taking an active part in the events making up the story.

Lubbock sees in these variants of viewpoint so many *successive stages of dramatising the story*, and with it, of improving its quality. This view does not hold.

Not only as shown above Sect. 19.5, is it unjustified to believe that the virtue of fiction is to be as much like drama as possible, but a close examination of novel structure shows a more or less continual change-over from one "stage" to another, not in the order as recommended by Lubbock, but quite randomly, and at such short intervals as to defy the attempt to see any of Lubbock's three stages as being predominant, let alone the only one present. The quality of a novel is not a function of the *type* of viewpoint, whether it be outside or inside the story, but of the extent to which the element of *change-over* from one type to another is employed—duality in action or dynamic duality. The quicker and the more natural such change-over, the more complete will be the canvas of life, and with it, the greater the involvement of the reader.

According to Lubbock, it was in Henry James's novels that the impersonal, objective method of fiction—the elimination of the author—had reached its peak. However, what the adherent of the "immanent" method would describe as a peak in quality of the craft of fiction does not turn out to be one in practice. A recent analysis of the appeal of fiction through "involvement" in the *Times Literary Supplement* (Oct. 21st, 1957) arrives at these conclusions:

"In wishing to pack every one of his scenes with the highest possible pressure of implication, James, great artist as he is, notoriously creates a world in which everybody hangs on every slightest nuance of everybody else's gesture or word, which is highly artificial, which, in its lack of breathing-spaces for stupidity, is very remote indeed from the world of 'real-life'. In Dostoevsky's work, on the other hand, the gaps, the uncertainties, the things reported, as it were, doubtfully or at second-hand are more like our own experience of life than the selected pregnant presentations of, say, 'The Awkward Age' are. The subtle explorations of so many modern novelists ... often seem to lead us in the end, with wit and urbanity, to a desolating sense of dryness at the heart of things, a frigidity or impotence in human emotions, a final inner emptiness. It is wholesome, sometimes, to return to the great writers of the past who could strike a spring of compassion from the human rock."

19.8 PLURALITY OF CODES AS A DEVICE OF THE CRAFT OF FICTION

The remedy for the growing estrangement between the code symbols and the world of real things is to use a number of codes, each of which is specially suitable for describing a particular aspect of the world. In this way, a more complete and more adequate picture is achieved than by using a single code, however perfect, and the "thinness of truly semiologic codes" thus becomes the "saving grace" of coded description. It necessitates bridging the gulf between the different codes, which has the effect of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

making us comprehend the manifold as such, as different from its coded description. As Lubbock expresses it: "And so by accepting and using what looked like a mere disability in the method, you convert it into a powerful and valuable arm with a keen effect of its own."

The concept of the inadequacy of truly semiologic codes, and of linguistic duality as its remedy, makes us better appreciate why the novelist uses the various devices of his craft, such as the panoramic, the scenic, the "framing" method which dramatises the author, and the various character viewpoints. *They are so many codes in which to describe the manifold*, each of them being particularly suitable for bringing out certain aspects, and their integration in the reader's mind producing the required complete picture of life.

It is true, in speaking of the novelist's code, we use the term in a sense somewhat different from that when speaking of language codes in Vol. I, Part IV. There we understood by it the system of ultimate (phonemes) or pen-ultimate (morphemes) linguistic symbols, or of words as units. Here we understand by it the novelist's devices for showing life from different individual angles and with different degrees of "concentration" or emphasis. At the same time, these codes are meant to describe the individual in whose mind the world appears, and the author may either tell the story, without taking part in it himself, and thus as if he were omniscient; or he may dramatise himself as narrator, either after the "frame" fashion or biographically.

19.9 SUMMARY

Thus, as observed above, our theory of language provides a suitable nomenclature for the "craft of fiction". The "dramatic", "scenic", "narrative", "pictorial" ways of presentation, the "author himself", and the "author's creations" in the frame of a story, and the "characters" are so many different codes, each presenting life from a different angle. The need for these arises through the inadequacy of any single such code to do justice to the complexity of the social life of our species. Their interplay is brought about by the duality of vision, which is inherent in the descriptions of the same complex from different angles. The greater the contrast in spite of the identity of the object described, the better will the difference between the characters be brought out.

However, the advantage in applying our theory of language to the craft of fiction lies not only in providing clearly defined concepts for the hitherto only descriptive terms as used by Lubbock. It makes us better appreciate certain devices whose use is either deprecated as unsuitable or recommended as vital, for novel structure, and the small advantage to be derived from others, though they are by some considered to be of great value. Thus Lubbock, in spite of appreciating the pictorial element, seems to regard it as a great merit of the book if the author, and thus the story-teller, is in the story himself: "Nothing is now imported into the story from without." But this method, if carried to the bitter end, would simply result in the novel becoming drama, whereas it is the virtue of the novel to remain a novel. If Lubbock had realised that the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

pictorial code, no matter whether the story-teller is within or without the story, is essential for an adequate description of the great canvas of social life, he would not have regarded dramatisation as so great an improvement of novel structure. In the novel, dramatisation is *not* a requirement for its own sake.

Since, as has been shown above, the dramatist does not use exclusively the linguistic code, but action is developed by things outside the coding devices, it follows that dramatisation cannot be the most important or most desirable device of the novelist who has to depend entirely upon the coding devices of language. Thus, the concept of 'code', if used sensibly, will make us beware of popular mistakes in criticsm.

19.10 ILLUSTRATIVE MATERIAL

The power of the analytical method reveals itself still more when it comes to considering critically the whole structure of a novel, not only its method.

19.10.1 Tolstoy, "War and Peace" and "The Kreutzer Sonata"

Lubbock, and many others, have criticised Tolstoy's *War and Peace* for being two novels instead of one. In addition to the personal story of its characters, there is the history of the Napoleonic invasion of Russia, and, as Lubbock expresses it, in the former "the war and the peace are episodic, the historical scenes being used as a foil and a background; in the latter, the personal history of the characters is a mere pebble on the beach". But Lubbock goes further; he believes that Tolstoy was "apparently quite unconscious that he was writing two novels at once". The events, he feels, are "inadvertently mixed and entangled". "War and Peace is like an Iliad, the story of certain men, and an Aeneid, the story of a nation, compressed into one book by a man who never so much as noticed that he was Homer and Vergil by turns."

The impossibility of such a view, and the contradiction with Lubbock's own appreciation of *War and Peace* as the greatest novel of all time, are unmistakable indications that Lubbock's view about the "two novels" is wrong, and that he has failed to understand the structure of that novel.

Let us now present that structure in the light of our theory of duality. In the story of Peace, that is, of the principal characters, their loves, ambitions, and hates are broken into from time to time by the great historical events. On the other hand, in the story of the War, whose central characters are Napoleon, Alexander, Kutusov, Murat, the story of the loves and ambitions of the young generation are relegated to the fringe. Thus we have a genuine duality of presentation, the private life or individual code describing events from an angle diametrically opposed to that from which the public life or national code sees them. This is anything but inadvertent mixing of two novels, or confusion of two designs. The *incongruity is fundamental* in that it makes us see the war from the viewpoint of the individual and the individual from that of war. How else could the fate of our species in times like that be adequately described? The device used by Tolstoy in "writing two novels" is that of duality in its

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

most pronounced form. Once this is understood, our criticism of Tolstoy's masterpiece is no longer exposed to such popular pitfalls as that through which Lubbock came to grief.

Tolstoy's novel *The Kreutzer Sonata* is a remarkable instance of the principle of duality in literary structure.

Tolstoy starts with the description of the company in the railway compartment, introducing himself as one of them. But soon, one figure of that "canvas", Pozdnyshev, becomes the story-teller, and thus the point of view from which the events are seen. Since Tolstoy himself is the recipient of the story as told by Pozdnyshev, his, the original author's, mind becomes the canvas upon which the story is projected.

In his general remarks, however, the "hero" ceases to be Pozdnyshev and it is again Tolstoy himself who paints for us the canvas of life and morals of contemporary society. Insofar as we are compelled to view our own condition through his glasses, the viewpoint changes once more, this time from the author to the reader.

The shifting of viewpoint is not to be understood as if the former point of view was done with the moment the new one was adopted. Closer examination of novel structure shows the subtle continual interplay of the different points of view, and that it is just the use of this device of the craft of fiction which makes a novel fascinating: through it the writer succeeds in unfolding in all its complexity "the painted veil which those who live call life".

19.10.2 Theodor Storm, "Der Schimmelreiter"

The author introduces first as the "canvas" upon which the story is painted the literary magazine *Hamburger Pappes Lesefrüchte* which he has read when a little boy, sitting at the feet of his grandmother. Its author gives a panoramic view of the seascape with himself riding on the dyke in very dirty weather, the shadowy figure of the "Schimmelreiter" brushing past him. This is the canvas as seen from the author's viewpoint.

Arrived at an inn, and accommodated among the dyke officials, the viewpoint shifts to that of the schoolmaster who begins telling the story of the "Schimmelreiter". The background is of course, again the seascape and especially the dyke, but now as it depicts itself in the mind of the schoolmaster. Thus, the schoolmaster, who is a part of the original canvas—the assemblage of dyke officials in the inn—becomes the point of view and the author's mind becomes the canvas upon which the story is depicted.

While the story is being told, the viewpoint changes repeatedly between that of the enlightened story-teller and that of the superstitious inhabitants of the marsh, that is to parts of the former canvas again. But more important is the change from the viewpoint of the story-teller to that of the main character in the story, namely the "Schimmelreiter". Not that the "Schimmelreiter" himself is made to tell the story, but by and by we begin to see the marsh and the conditions of life there through the eyes of the "Schimmelreiter", and begin to judge the relations of man and nature in these

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

parts with his clear head and proud, unflinching soul. Again a part of the canvas has become the point from which to see humanity, and among it people of the type of the schoolmaster and especially the superstitious inhabitants, the representatives of the previous viewpoints.

And so it goes on, viewpoint and canvas, or point and plane of content, changing places in the story, until we realise that a more fundamental change has taken place when the reader, who thought of the story as just a cinematic picture, something he could coolly and detachedly behold from "above", in more senses than one, realises that it is his own life and character which are here examined and subjected to a truly last-judgement scrutiny: "what would the 'Schimmelreiter' think of me?" Thus it is we, our own souls, who have now become the canvas, the "Picture of Dorian Gray".

Involvement, without which there can be no true literary impression, proceeds in stages each being the dual of the other, part of the former canvas changing into a viewpoint and the former viewpoint becoming part of the new canvas. The transitions are unaccountable; they constitute what we call the Craft of Fiction.

19.10.3 Goethe, "Iphigenie" and "West-Östlicher Divan"

Goethe's *Iphigenie* is permeated with duality.

There is the opposition of Iphigenie in Aulis and Iphigenie in Tauris, the first representing the occasion when Iphigenie almost becomes the victim of Diana, the second when her brother Orestes is in the same position.

The earthly relationship of brother and sister between Orestes and Iphigenie is duplicated by its divine counterpart between Apollo and Diana. And there is the cross-relation: the wish to placate Diana was the cause of the threatened sacrifice of Iphigenie, and the wish to please Apollo is the cause of Orestes' implication and threatened sacrifice.

The plot itself is a perfect duality: Apollo's dictum promising Orestes recovery from disease if he brought home the sister from Tauris which makes Orestes embark on the journey to Tauris, is understood by Orestes as referring to the image of the heavenly sister, Diana, but is found to imply his own living sister, Iphigenie.

The poet's device of changing the viewpoint is nowhere used with greater virtuosity and spontaneity than in Goethe's *West-Östlicher Divan.*

In the poems where Goethe speaks through, and as, Hatem, the viewpoint is changed from that of the author to that of one of his creations. The same is true where Goethe speaks through Suleika. But in the poem *Hatem* a most significant shift of viewpoint is performed in a way quite singular in literature: in verse three, the word required to rhyme with "Morgenröte" is unexpectedly "Hatem", for which we instantly substitute "Goethe". This shifts the viewpoint back to the author, and the Hatem-Suleika relation is revealed as that between Goethe and Marianne v. Willemer.

Moreover, Marianne herself has written poems which are the perfect duals of Goethe's, and as such are incorporated in the *West-Östlicher Divan.*

What makes the *West-Östlicher Divan* so remarkable for us is not so much the use

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of the duality device, since there cannot be any truly great literary work without it, but the *conscious* use of it and the pleasure the poet takes in exhibiting his method.

19.11 THE LINGUISTIC DUALITY OF POETIC DICTION

The power of linguistic duplication or duality is, perhaps, nowhere better exemplified than in poetic diction, more particularly on the effect of the prosodic tools of rhyme and metre.

Both are fetters of thought, and at the first glance it seems like treason against the rules of logic to subject the expression of thought to compulsion, perhaps even slightly distorting, by such formal factors, and seemingly to no other purpose but that the same sound should be heard again after a specified number of syllables. Without at least a modicum of such compulsion hardly any verse could be said to have arisen, which, incidentally, is one of the reasons why poetry in a foreign language is always harder to understand than prose. But, in spite of such "prosaic" criticism of verse, poetry has at all times and with all nations exercised great power over the human mind through just those "thought-distorting" features which are extraneous to pure reason.

What is the cause of this? We start with the observation that a good verse through its emphatic effect produces the impression that *the thought was already pre-destined in the language*, and all the poet had to do was to find it out (Schopenhauer). But this means that rhyme and metre act as substitutes for logical proof or, at least, the considerations of reason. They thus represent the Eigencontrol (see Sect. 21.4) which language as form exercises over "expression", and which is comparable to that which in prose reason and logic exercise over "content". If we restrict the term "language" to mean "expression" as different from "content" we can say that rhyme and metre, and indeed other features of linguistic structure, provide the "inner" or Eigencontrol of thought, whereas that exercised by logical reasoning is a sort of "outer" control since it refers to "content". Every statement we make requires, if not exactly proof in the strict sense of the term, so yet sufficient reasons, which it is the function of the intellect to provide; in poetry, statements are not proved, but receive an equally powerful support by attestation through rhyme and metre.

We can, therefore, say that in poetic diction control of thought is provided by the effect of its very tools, rhyme and metre, — to name only these two—which create the impression that the ideas as expressed by the poet were already pre-formed in the language, and thus, appear attested by language itself. Thus, formal agreement of structural features takes the place of logical agreement. Accordingly, Karl Kraus has said of rhyme "Er ist das Ufer wo sie landen, sind zwei Gedanken einverstanden", which can be translated as, "It is the shore on which they land, if thoughts each other understand."

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

FIFTH PART

THE CALCULUS OF LINGUISTIC OBSERVATIONS

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XX

MATHEMATICAL CHARACTERISTICS OF LANGUAGE STRUCTURE

It is a consequence of the many-sided nature of language that an appropriate calculus of linguistic observations should be a complex net-work of different strands of statistics. In fact, three different types of statistics are required for dealing adequately with language problems.

The methods used so far in stylostatistics and statistical linguistics belong to the branch of *number statistics* which, by and large, is what one usually understands by the term "statistics" (as used in economics, demography, biology and medicine). After its originators and most prominent representatives in more recent times, it can be called Laplace-Gaussian and Yule-Pearson-Fisher statistics, respectively.

Information theory uses the branch of statistics known as *physical statistics* or, better, statistical physics, more widely known under the name of its originators as Maxwell-Boltzmann statistics. Although the methods of the two types of statistics and the parameters for characterising statistical masses are rather different, yet the probability basis is identical, and both Laplace-Gaussian and Maxwell-Boltzmann statistics are sometimes referred to as *classical statistics*. It is mainly the fields of application—the masses of denumerable events, mostly of social life, in number statistics and the distribution of energy in statistical physics—which make the two branches of classical statistics different.

So far it was only these two branches that were applied to language problems, and this may account for some of the criticism, if not aversion, which statistical methods sometimes encounter among linguists, namely that linguistic expression was too much a matter of conscious choice for its results to be amenable to pure chance considerations. In spite of the many misunderstandings in this respect (see Section 1.7), there is an element of truth in it. The types of statistics used so far, the Laplace-Gaussian and the Maxwell-Boltzmann variety, are suitable for dealing with characteristics of *expression*, but they do not reach *content*; even where the methods deal with the frequency of use of particular vocabulary items, it is only a formal characteristic of content—frequency of occurrence—which is investigated as the variable, not content itself.

But the criticism is wrong in ascribing the insufficiency of statistical linguistics to statistics in general: all one could say is that the methods used *so far* are not quite

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

suitable when it comes to dealing with content. The probability models which are characteristic for the two branches of classical statistics need not be universal and, in particular, need not be appropriate for dealing with content. The case is not only analogous to the need for conceiving a new type of statistics known as the Bose-Einstein statistics for Quantum physics when the classical forms were found not to be in agreement with the experimental observations, but, as will be shown, it is precisely the Bose-Einstein type of statistics in the form of the Random Partitioning function, which appears appropriate when dealing with language content.

The step from the statistics of information theory to those of structured meaning is analogous to that from stylostatistics to statistical physics as used in information theory, but even more drastic, since the "New Statistics" are Quantum statistics and, as such, different from classical statistics even in their probability foundation.

20.1 AN EXORDIUM ON LINGUISTIC SAMPLING

20.1.1 Equivalence of the proportions obtained by sampling counts and the proportions in the whole of a literary text

It is the particular geometrical properties of linguistic expression which determine the methods of linguistic statistics, and make it possible to put linguistic sampling on a scientific basis. This is directly evident for the written language, and insofar as the latter is essentially, and necessarily so, of the same structure as the spoken language, it must also be true for the latter. But whatever views one may hold about the controversial matter of similarity of structure in the spoken and the written language, the application of a scientific theory of sampling requires the transcription of the spoken language in whatever sort of symbols which makes it, for the purpose of sampling, again a written language.

The relation between type, token, text in the written language being essentially that between points, lines, planes of geometry, the sampling procedures for collecting information about the great mass of linguistic forms in literary texts can be based upon *geometrical probability*.

1. *Point sums as estimators of relative areas*

In Figure 14 a small irregular area B is enclosed in a larger irregular area $(B + W)$. The probability that the point located at random in $(B + W)$ will also lie in B is, by definition,

$$p = \frac{A_B}{A_{(B\ +\ W)}}$$

the ratio of two areas.

The expected number of points $S(X)$ which fall in B in a particular sample count containing altogether n points is

$$E\left[S(X_b)\right] = np = n\frac{A_B}{A_{(B+W)}} \tag{1}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

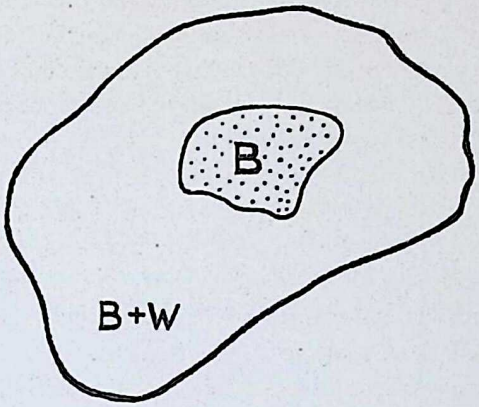


FIG. 14. (F. Chayes, 1956)

As a proportion, μ, of the total count of n points, this is

$$\mu = \frac{1}{n} E\left[S(X_b)\right] = p = \frac{A_B}{A_B\,(+\,W)} \tag{2}$$

Since its expected value is the ratio of the smaller to the larger area, the proportion of the total sample count that falls in the smaller area is an unbiased estimate of that ratio.

For language material, disregarding the space between the lines of literary text and letting Ty (Type) stand for letter (also for space between words and for punctuation), and To (Token) for occurrence within lines, the ratio of the number of a particular sort of Ty, say, the letter *e*, chosen at random on the page to the total sample count of To, is an estimate of the ratio of the area occupied by the letter *e* to the total area (all letters) of the page, and thus an estimate of the probability of *e* in a page of the text.

Having made a phonemic transcription of the text and letting Ty stand for phonemes, an estimate of the probability of a particular phoneme in a page of text is obtained in the same way, viz. as the ratio of the number of that phoneme in a sample count to the total number in the count.

Denoting by "point" not ultimate units, letters or phonemes, but syllables or words, the frequency of occurrence of a particular word, say, in a count of n words occurring on a page, is an estimate of the probability of that word in a page of the given text.

2. *Parallel lines as estimators of relative areas* (*F. Chayes, 1956*)

However, counts of linguistic forms are not usually made at random on a page—which is more difficult than it seems—but along the lines of text, and it is now incumbent upon us to show that this, too, leads to an estimate of the relative area occupied by a particular linguistic form.

The areas under the curves in Figure 15 are obviously

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

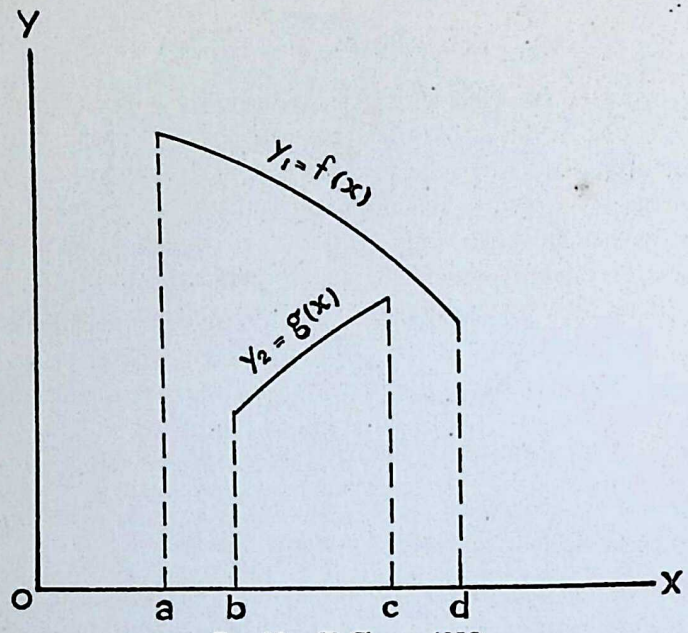


FIG. 15. (F. Chayes, 1956)

$$A_1 = \int_a^d y_1 \, dx,$$

$$A_2 = \int_b^c y_2 \, dx$$

Let us suppose that ordinates are to be erected at points along OX chosen simply at random in the region $a < x < d$. The element of frequency is thus dx, and the total frequency is:

$$F = \frac{1}{d-a} \int_a^d dx = 1 \tag{3}$$

The expected value of y_1 is then

$$E(y_1) = \frac{1}{d-a} \int_a^d y_1 \, dx = \frac{A_1}{d-a} \tag{4}$$

and for y_2

$$E(y_2) = \frac{1}{d-a} \int_b^c y_2 \, dx = \frac{A_2}{a-d} \tag{5}$$

(In the regions $x < b$ and $x > c$)

The ratio of the average ordinates is a consistent estimate of the ratio of the *areas*, for it is an estimate of the parent value μ_2/μ_1, and

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\frac{\mu_2}{\mu_1} = \frac{E(y_2)}{E(y_1)} = \frac{A_2}{(d-a)} \cdot \frac{(d-a)}{A_1} = \frac{A_2}{A_1} \tag{6}$$

In terms of Ty and To standing for linguistic form (e.g. letter, phoneme, word) and its occurrence along the lines of a text respectively, this means that the ratio of the average number of the particular form per line of text relative to the length of the line is an estimate of the relative area occupied by that form on a page of text, and the ratio of the average numbers of two particular forms per line of text an estimate of the ratio of the corresponding areas of the page occupied by these two forms. Multiplying the averages by N, the number of lines examined, we obtain:

$$\frac{N\mu_2}{N\mu_1} = \frac{A_2}{A_1} \tag{6a}$$

and can thus say that the ratio of the *sums* of occurrences of the two linguistic forms are an estimate of the ratio of areas occupied by them.

3. *The area-volume relation*

Ultimately, we are not interested in the probabilities per page, but in those for the whole book, the text. Considering that a book is a three-dimensional body — which is, of course, true also for the page if its thickness is taken into account, which, however, does not concern us, the printed page being, for the purpose of linguistic sampling a two-dimensional entity—, we must try to extend these considerations so as to arrive at estimates of the proportions in the text.

If the area of a section of a solid parallel to the xy plane is a function of z, $A = f(z)$, and sections through the solid can be chosen simply at random normal to OZ in the region $c \leqslant z \leqslant d$, the element of frequency is dz, the total frequency is

$$F = \frac{1}{d-c}\int_c^d dz = 1 \tag{7}$$

and the expected value of A is

$$E(A) = \frac{1}{d-c}\int_c^d A dz = \frac{V}{d-c} \tag{8}$$

where A and V represent area and volume respectively. We have at once that

$$\frac{E(A_a)}{E(A_t)} = \frac{V_a}{V_t} \tag{9}$$

where V_t is the total volume concerned and V_a the part of it occupied by a particular component. *Thus the ratio of the area occupied by component A to the total measurement area is a consistent estimate of the volume percentage of component A in the total body.*

Since the relative frequency of linguistic points (Ty) provides an unbiased estimate of the area per page (To), and since the latter provides an unbiased estimate of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

volume occupied by a particular Ty as part of the text (Te), it is true to say that *the relative frequency of a particular linguistic form in a sample count is a consistent estimate of the volume percentage occupied by that form and, thus, of its probability in the text.*

To summarise, we have seen that linguistic counts made in the usual fashion are consistent estimates for *areal* proportions, i.e. in the pages of a literary text, and that the "composition" of an area selected at random is a consistent sample of *volumetric* proportions, i.e. the proportions in the whole of a literary text.

20.1.2 Reproducibility of estimates of proportions of linguistic forms

In Sect. 20.1.1 it was argued that the probability p of a point chosen at random in the area $(B + W)$ lying also in B is equal to $A_B/(A_B + A_W)$. Since it must fall in W if it did not fall in B, the probability of its falling in W and not falling in B is

$$q = \frac{A_W}{A_B + A_W} = 1 - p \tag{10}$$

We may say then that we are "observing" a quantity, X, which has the value 1 whenever a point lies in B, and the value 0 whenever a point does not lie in B. The frequency distribution of X is

X	0	1
$f(X)$	q	p

The population mean will be

$$\mu = \Sigma X f(x) = (0)(q) + (1)(p) = p \tag{11}$$

If the points are "drawn" in batches of n, we record only the sum of the X's, $S(X)$. But obviously

$$\mu_{S(X)} = np \tag{12}$$

and if we reduce the result to a percentage basis we have since $\bar{X} = (1/n)\, S(X)$

$$\mu_{\bar{x}} = \frac{1}{n}\mu_{S(x)} = p \tag{13}$$

Whether we draw the points in batches of the same or different size, each drawing estimates the same population mean. However, the variance of the estimate varies inversely with the number of points n on which it is based. From the frequency distribution we can readily calculate the population variance, or variance attaching to an estimate based on batches of size $n = 1$.

By definition, this is

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\begin{aligned}\sigma_1^2 &= \Sigma(X-\mu)^2 f(X) \\ &= (0-p)^2 q + (1-p)^2 p \\ &= pq\end{aligned} \tag{14}$$

Accordingly, for $n > 1$,

$$\sigma_n{}^2 = \frac{1}{n}\sigma_1{}^2 = \frac{pq}{n} \tag{15}$$

and the standard deviation of an estimate of p based on n points is, in percentage of the whole,

$$\sigma_n = 100\sqrt{\frac{pq}{n}} \tag{16}$$

From equation (16) it is clear that, for fixed n, σ is symmetrical about a maximum $p = 0.5$, and that for a given p, σ varies inversely as $\sqrt{n}$. However, it is of far greater importance that for reasonably large n, σ is rather insensitive to even large differences in p and fairly sizeable differences in n.

The test function used to decide whether the relative frequency of a particular linguistic form is reproducible by word counts carried out on different pages of the same text, is the *Bernoullian law* (Vol. I, Sect. 5.5). This law enables us to connect deviations from the population mean in terms of the standard deviation with probability of occurrence, and deviations exceeding three times the standard deviation are so unlikely to occur on pure chance that we neglect the possibility in order to arrive at practical conclusions.

20.1.3 Illustrations from stylostatistics

To illustrate the use of the formulae given in Vol. I, Sections 17.2.1. and 17.2.3 in matters of stylostatistics we have chosen the problem of the passage in Mark 16, 9-20, which Bible scholars have reason to believe not to have been written by the same person as the rest of the Gospel. It comprises 167 occurrences with a vocabulary of 92 different words.

Selecting now certain of these words, we may be interested first in the question of whether their occurrences in Mark 16, 9-20, and in other specified parts of the New

TABLE 78

Word	N.T.	Mt.	Mk.	Mk. 16, (9-20)	Lk.	Jo.	Paul
Μαρία	54	11	8	1	17	15	1
ὅτι	1285	141	101	1	173	271	282
εἰς	1753	216	167	3	223	182	426
Text length	137328	18278	11229	167	19404	15420	32303

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Testament was such as to be compatible with the view of their occurrence being random samples of the respective basic probabilities of these words in the New Testament.

The number of occurrences of our three words in each of the six parts of the New Testament under consideration can be calculated under the assumption of random sampling from the whole of the New Testament, according to the formula

$$a = np$$

with the standard deviation $\sigma^2 = \gamma_1^2 p(1-p)$, (17)

where[1] $p = m_0/m$ (18)

and $\gamma_1^2 = n\dfrac{m-n}{m-1}$ (19)

The quantities required for the calculation are listed below.

TABLE 79

Word	New Testament	
	p	$p(1-p)$
Μαρία	.00039	.00039
ὅτι	.00933	.00924
εἰς	.01272	.01256

	Mt.	Mk.	Mk. 16 (9-20)	Lk.	Jo.	Paul
γ_1^2	15792.19	10308.22	166.67	16648.63	14248.08	24679.50

The inspection of the quantities is very instructive. On the one hand, only for very small p ($< .01$) will $1-p$ be sensibly equal to unity, and $\sigma^2 = np$, and thus equal to a which is a characteristic of the Poisson distribution. On the other hand, the greater the sample size, the more will the factor $\dfrac{m-n}{m-1}$ differ from unity and the formula for σ^2 differ from that for the Bernoullian distribution: $np(p-1)$. It follows that the greater n compared with m, the less reliable will be the use of the formulae based upon the assumption of the applicability of the Bernouillan law as the limiting case, and the more advisable to use the formulae derived by the hypergeometric function, Vol. I, p. 316, which take the size of the factor $\dfrac{m-n}{m-1}$ duly into account.

The expected values, together with their standard deviation, are listed in table 80.

[1] m = text length, m_0 = frequency of particular word, n = sample size.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

TABLE 80

Word	α					
	Mt.	Mk.	Mk. 16 (9-20)	Lk.	Jo.	Paul
Μαρία	7.13	4.38	.065	7.57	6.0	12.60
ὅτι	170.53	104.76	1.56	181.04	143.87	301.39
εἰς	252.51	142.84	2.12	246.82	196.14	410.90
Word	σ^2					
	Mt.	Mk.	Mk. 16 (9-20)	Lk.	Jo.	Paul
Μαρία	10.16	4.02	.065	6.49	5.56	9.63
ὅτι	145.92	92.25	1.54	179.29	131.65	298.48
εἰς	198.35	129.47	2.09	209.11	178.96	309.98
Word	σ					
	Mt.	Mk.	Mk. 16 (9-20)	Lk.	Jo.	Paul
Μαρία	2.48	2.01	.255	2.55	2.36	9.10
ὅτι	12.08	9.76	1.24	13.38	11.48	17.28
εἰς	14.087	11.38	1.45	14.50	13.38	17.61

Inspection of the results shows that most of the observed values for word occurrences lie within the range $\alpha \pm 3\sigma$, except the occurrence frequencies of Μαρία in Mark 16, 9-20, Luke and John, that of ὅτι in John, and εἰς in Mark. This means that apart from the cases mentioned, the observed occurrence frequencies are such as to be compatible with the assumption of their being random samples from the respective probabilities in the total.

But the main problem in which we are interested is whether the occurrence of the vocabulary items in Mark 16, 9-20 was such as to justify the conclusion that by the token of use of vocabulary it could not be regarded as just a random sample from the Gospel according to Mark. To answer that question we may proceed in different ways. We may regard Mark as the population and estimate from it the frequency of certain words in a sample of the size of Mark 16, 9-20. Or we may start from Mark 16, 9-20 as the sample and estimate the frequency of a particular word in the whole Gospel as the population. However, neither of these methods is theoretically quite satisfactory in this case because the size of Mark does not warrant the assumption of its being the population. The test applied below under that assumption should therefore be regarded only as an illustration of the testing procedure as such.

The third possibility, and the one which is more in accordance with the fact of the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

number of words in Mark being finite (11,229) is to test whether the frequency of occurrence of a particular word in each, Mark 16, 9-20 and in the rest of the Gospel was compatible with regarding both as random samples from a larger whole.

Regarding first the Gospel as the population, we can estimate by formula 25, Chpt. 16 Vol. I, the expected number of occurrences of a particular word in the passage in which we are interested (Mark 16, 9-20). We now select words of particular interest e.g.

TABLE 81

Word	Mk.	Mk. 16 (9-20)
λορενδείσα	3	3
ἑτέρα	1	1

For λορενδείσα we have by formula 25, Chpt. 16, Vol. I.

$$\alpha = 167\,\frac{3}{11229} = .0434$$

$$\sigma^2 = 167\,\frac{11062}{11228}\,\frac{3}{11229}\left(1 - \frac{3}{11229}\right) = .0430$$

$$\sigma = .2076$$

We observe the sensible equality of α and σ^2, which shows the distribution as being governed by the law of rare events. Since $\alpha + 3\sigma$ remains far below the observed frequency of the word in Mark 16, 9-20, it would support the conclusion of the difference in style, as far as the use of λορενδείσα is concerned.

For ἑτέρα we have

$$\alpha = 167\,\frac{1}{11229} = .0134$$

$$\sigma^2 = 167\,\frac{11062}{11228}\,\frac{1}{11229}\left(1 - \frac{1}{11229}\right) = .0132$$

$$\sigma = .15$$

The conclusion is the same as before.

If we regard the passage in Mark 16, 9-20 and the rest of the Gospel as samples from a larger whole, we calculate the statistics as follows (Vol. I, p. 328)

$$\alpha = np^*;\quad \sigma^2 = y_{III}^2 p^*(1 - p^*), \tag{20}$$

where

$$p^* = \frac{m_0 + 1}{m + 2},\quad y_{III}^2 = \frac{n(m + n + 2)}{m + 3} \tag{21}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

For λορενδείσα ($m = 11062$, $m_0 = 3$, $n = 167$) we have

$$\alpha = 167\,\frac{4}{11064} = .060$$

$$\sigma^2 = \frac{11231 \times 167}{4065}\,\frac{4}{11064}\left(1 - \frac{4}{11064}\right) = .061$$

$$\sigma = .247$$

For ἑτέρα ($m = 11062$, $m_0 = 1$, $n = 167$)

$$\alpha = 167\,\frac{1}{11064} = .015$$

$$\sigma^2 = \frac{11231 \times 167}{11065}\cdot\frac{2}{11064}\left(1 - \frac{2}{11064}\right) = .0305$$

$$\sigma = .175$$

Since $\alpha + 3\sigma$ is much below the observed number of occurrences of these words in the passage in question, the style difference, as regards the use of these words, between the passage under consideration and the rest of the Gospel can be regarded as statistically established.

20.1.4 Sampling efficiency of linguistic material

The argument so far was based on geometrical probability and concerned only with the fact that relative frequencies of linguistic forms obtained from sample pages of a linguistic text can provide consistent estimates of the probability of such forms in the whole text. *How* to take samples from a linguistic text in order that the principles of geometrical probability should give the desired results, and *how large* to take them, depends very much upon the type of linguistic form. In general, the less the occurrences of Type for which we sample is connected with the content of the text, the less need we trouble about superimposing a random sampling scheme upon the natural occurrence of the Type in the text, for the simple reason that it will, by and large, be already random. This is, for instance, true for letters and phonemes (Sect. 20.2.1).

Considering the geometrical nature of the Type-Token-Text relation, it is not surprising that the theory of sampling for the purpose of linguistic analysis should conform closely with that of another similar geometrical configuration, viz. that required for the modal analysis of rocks. A rock is a mineral aggregate and the composition of the rock expressed in terms of the relative amounts of minerals present is called a mode. The method as such is then referred to as Modal or Thin Section Analysis. It consists of scanning thin sections of rock for the various minerals present. The "points" of Modal Analysis are the grains of the different types of mineral as seen under the microscope in the case of fine-grained rocks on the surface of a thin section, the "plane". The analysis consists in determining the proportion of a particular mineral

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

in the rock, that is in the three-dimensional configuration, on the basis of the proportion occupied by it in the thin section, the area, which again is estimated from line segments occupied by the particular mineral. Of great practical interest for the research in quantitative linguistics is the fact that Modal Analysis has developed mechanical time-saving counting procedures which, with certain modifications, could be made available for linguistic counts.

All this shows how closely akin is the "nature of the book" to the "book of nature" whose pages, as far as rocks are concerned, are the thin sections of rocks.

20.2 THE UNIVERSES AND POPULATION PROBABILITIES OF LITERARY STATISTICS

20.2.1 The universes of literary statistics—The universe of content

We have seen that phonemic (alphabetic) and quantitative characteristics (word length in terms of letter, phoneme or syllable number) of language are, on the whole, independent of individual style, and thus language-conditioned, whereas qualitative characteristics such as vocabulary, and sentence length are to a great extent style-conditioned, with grammar characteristics occupying a medium position while inclining more strongly to being independent of style, and thus to being language-conditioned.

This has important consequences for the sampling of linguistic data. The more language-conditioned a linguistic characteristic is, the smaller may be the sample which can be considered representative of the statistical population. It follows that reliable information about quantitative characteristics may be obtained from even small samples; when sampling for grammar characteristics, we shall have to be careful not to include stylistic outsiders if we wish to get information about the average use of grammar forms by the members of the speech community. This means that we have to extend the field of observation so as to be able to distinguish normal variations in the use of grammar forms from extraordinary. When it comes to sampling for the use of vocabulary, the situation is altogether different. There is no such thing as an average vocabulary, and a representative sample of vocabulary can only mean the integration of different strata of vocabulary from a very great number of different texts (written or spoken language). It follows that sampling for vocabulary use requires not only greater samples, but also great discretion and judgment. This justifies Trubetzkoy's warning.

The differences in the method of sampling for the three types of style characteristics, quantitative (phonemic), grammar and qualitative, correspond to differences in what is called the statistical population of the characteristics, or the universe of discourse.

For truly quantitative style characteristics and for phonemes (letters) it was found that even small samples (of say 1000 letters or phonemes) appear to be representative of the statistical population, provided that due care is taken to spread the sample in a random way over a great part of the text. We say in this case that *the universe is one of simple sampling*, and mean by it that a specified linguistic form, say a specified

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

phoneme, has a constant probability, independent of the text samples. It is a comparatively simple matter to ascertain these quantities, and that is why the universe is called one of "simple sampling".

Grammar style characteristics, although more closely related to quantitative characteristics than to vocabulary, may show significant variations with individual style, if the field of observation is sufficiently extended. This means that we cannot speak of the probabilities of such characteristics as strictly constant throughout the universe, though some of them can still be considered as constant (e.g. nouns, adjectives and pronouns), to a greater extent than others (e.g. verbs). Since a reasonably small sample will no longer give an adequate picture of the universe of grammar forms, sampling must extend over a field large enough to include possible "outliers", i.e. writers having certain idiosyncracies in the use of grammar forms. In addition to the variation of constant probabilities due to fluctuations of random sampling, there is variation of these basic probabilities due to assignable causes. The variation which in the universe of simple sampling is governed by a simple law of chance, such as the Binomial law or the Poisson law, consists now of two parts: an aleatoric part governed by a law of chance and a "physical" part, representing the element of choice. We may speak here in contradistinction to the phonemic universe discussed above of the tagmemic universe. This type of universe is known in statistical theory as a *Lexian universe* (see Vol. I, sect. 6.4,18.6).

The element of choice assumes greater importance when sampling for vocabulary style characteristics. It is important to remember that language data are peculiar in this, that they show gradual transitions from simple sampling to sampling from which the limitations of simple sampling have been removed, and sampling for vocabulary characteristics is just one further step in that direction compared with sampling for grammar characteristics.

The choice element has become so powerful that we cannot any longer aim at separating it in a simple manner from the aleatoric effect, the chance element. We therefore do not aim at a description of the universe with the choice element eliminated, but rather at an integration of such styles which together might give an adequate picture of the use of vocabulary in the speech community. In place of the conventional parameters for describing universes like that of quantitative and grammar style characteristics (mean, standard deviation), we now have to use a complex parameter, the Characteristic or v_m as an overall index of the type-token relation in the universe. We may here speak of the *vocabulary universe of language*. A related type is what might be called the *taxonomic universe of content*.

Instead of aiming at a stratified sample from a great number of texts and writers, we may also regard each writer as a universe. This will be the correct procedure when individualisation of style has become pronounced.

20.2.2 The linguistic population probabilities

The basis for the application of statistical methods to linguistic data is the quantitative

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

interpretation of the "langue-parole" dichotomy (Sect. 1.3, 5.3, Vol. I). According to it, "la langue" is the total of the linguistic forms shared by the members of a speech community plus their probabilities of occurrence, and "la parole" represents random samples of that population.

The probabilities themselves were explained, using a concept of physics or chemistry, as the precipitation of the speech occurrences with time, which makes them appear a rather hypothetical quantity. For biological random variables, the differential rates of natural production of the various categories of the variable provide the population probabilities. Language is neither a purely natural product, nor is it completely artificial and the conceptions of population probabilities of biological and physical random variables do not apply as a matter of course.

Where do these linguistic probabilities reside and how are they connected with the variable? The answer to that question, rendering linguistics a safe anchorage for statistical theory, is provided by the discovery that these probabilities are the relative frequencies known as the *functional burdening* of specified linguistic quantities. The term "functional burdening" is here understood as referring not only to phonemes—where it was used originally—but to other linguistic units and characteristics, such as letters, word-length in terms of syllable, letter or phoneme number, morphemes, words, which means that they reside in the dictionary, and ultimately in the brains of the members of the speech community.

The conclusion is based upon empirical observations. In Chapter IX it is shown that the frequency of English phonemes in speech (and no doubt also in writing) can be explained as a random sampling effect from the statistical population of such phonemes if the functional burdening according to the dictionary is used as the population probabilities. In Chapter X, it is shown that the distribution of word-length according to phoneme and letter number in "la parole" can be understood as a derivative distribution (a moment distribution) of the distribution in the dictionary as the basic one.

With words as units the matter is more difficult since a dictionary count of the semantic burdening of words, i.e. their burdening with meaning, as a random sample of which we would like to explain the frequency distribution of occurrence in continued pieces of writing or speech, is not an easy matter. No reliable count of that sort is available, to the author's knowledge. We have, however, good reason to assume that such a relation exists. In Sect. 6.5, Vol. I, the remarkable degree of stability of frequency of use of words in different periods of time was interpreted as a pointer to similarity of the complex of occasions for the use of these words. And in Sect. 11.5, l.c. the same explanation was put forward for the similarity of word length in translation equivalents.

The interpretation of the functional burdening of linguistic symbols according to the dictionary as the population probabilities in the various linguistic universes transforms what was before a mere hypothesis, into a scientific theory.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

20.3 THE ERGODICITY OF WORD MASSES

1. It is a characteristic of linguistic distributions that their properties may be interpreted in terms of either number statistics or statistical physics (Vol. I, Sect. 1.5). According to the former, the properties of empirical linguistic distributions are derived from those of a hypothetical statistical universe of which the empirical distribution is conceived to be a random sample. According to the latter, the universe is not primary, but represents the limiting form to which the linguistic system will, under certain circumstances, tend. We describe below some of the essential features of linguistic systems, using both types of interpretation.

2. We start with the fundamental tenet of de Saussure, viz., that the signs of language are *independent* of their content or meaning. Now let us review the results of the quantitative study of language in the light of the definition of random sampling (as given by Kendall)[2] as a method of selecting specimens of an event which is *independent* of the event itself, or, briefly, as a method of sampling in which the units selected for examination are independent of what we are sampling *for*. Linguistic sampling, no matter how varied, is essentially one of words. If our purpose is to make a dictionary count or a text count of letters, of phonemes, of letter or phoneme groups, of word length in terms of letter or phoneme number or syllable number, the units selected directly are words, since the sample is given in terms of the number of words examined. We may, of course, sample in a different way, e.g. by counting the items in which we are interested in a specified length of the linear sequence of letters, say, plus spaces between words and punctuations, but nothing is gained by it, since, as we have seen, samples in terms of word number *are* random samples of the smaller units, such as letters, phonemes, word length, etc. Indeed, this forms an important characteristic of language statistics.

In the light of the definition of random sampling given above, we must conclude that what we select or sample *by*, namely words, is independent of what we sample *for*, be it the ultimate linguistic units such as letters and phonemes, or the secondary linguistic characteristics such as word length. Considering now that words have meaning, and in texts are arranged according to meaning, our conclusion is in full agreement with the fundamental tenet of de Saussure mentioned at the beginning of this paragraph, namely, that the signs of language are *independent* of their content or meaning.

The main result of statistical linguistics as described above, namely, the independence of the quantitative characteristics of language from content or meaning, is fundamental—though seldom recognised to be so—for the whole concept of information theory. Only because of that independence is it possible that the smaller code units can be manipulated with a view to reducing the time, and with it the cost, needed for transmitting a message, without interfering with the content of the message. We may express this by using the term "Ergodicity" which originated in statistical physics.

[2] M. G. Kendall et al., 1938.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

"Ergodicity" denotes a property of certain systems which develop through time according to probabilistic laws. Under certain circumstances a system will tend in probability to a limiting form which is independent of the initial position from which it started. This is the Ergodicity Property . It may be regarded as a form of the Law of Large Numbers when applied to stationary processes.

Applied to the linguistic distributions we have in mind, we may say that no matter what the content of the message, that is no matter what the initial conditions, as time goes on the masses of words used in the message will show a sensibly stable distribution of the smallest units, letters or phonemes, in any part of the message.

3. The same applies when we sample a text—not a dictionary—for frequency of occurrence of vocabulary items. No matter which part of the text we sample, or which text by a particular writer, the Characteristic K epitomising the vocabulary-occurrence, or type/token relation, is sensibly constant, which means that the frequency property of vocabulary is again independent of the content which necessarily differs somewhat between different parts of the text, and much more between different texts.

The above statement about independence of word length from content, although true to a large extent, is not so general as is that of the independence of letter or phoneme occurrence. Subject and style begin to exert some influence upon the use of words of different length. Such a qualification of the statement about independence is still more important in the case of frequency of occurrence of vocabulary items, indeed so much more that the Characteristic K was originally regarded as a style characteristic. However, with these provisos it is true to say that even word length and frequency of use of words—not particular ones, though—are also largely independent of content, and thus support de Saussure's tenet about the independence of sign and meaning. This does, of course, not mean that a particular word will always have the same frequency in different samples from a given text, let alone in different texts. It is only the *group* of vocabulary items occurring with a specified frequency which will represent approximately the same proportion of the total vocabulary in different samples or texts. Members of that group may be different vocabulary items, though a good many will be the same.

All this may seem to be more exciting to the statistician than to the linguist, until we realise what the fact that our word samples are random as regards the relative frequency of letters, phonemes, word length and vocabulary items with specified occurrence frequencies means for the use of words in speech and writing. Briefly, we may say that although speech requires primarily the choice of words to fit the thoughts we wish to express, yet its characteristic sound pattern, its speed, and with it that of thought, and its uniformity and diversity of expression are language-conditioned, and thus largely outside individual choice. Deviations from the language-conditioned pattern of sound, speed of thought, and uniformity and diversity of expression, are largely governed by chance, that is they are, as a rule, not greater than what might, with a high probability, occur on pure chance. This provides us with a standard of comparison for such language-conditioned characteristics of expression in speech and

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

writing against which to judge a deviation which is possibly in excess of what may still be reasonably attributed to chance.

4. In Section 3.3 the regularities in the distribution of vocabulary over different parts of a text were explained as the result of random partitioning of the total vocabulary. This means that we ascribe to the total frequency distribution the role of a statistical universe and derive the vocabulary distribution as the result of random partitioning that universe.

However, as observed above, it is a characteristic of language statistics that its properties may often be construed in terms of either number statistics or statistical physics since linguistic distributions partake of the properties of both. The stability of the distribution of vocabulary among different parts of a text can also be explained as an ergodic property.

No matter what the content (the initial conditions), the system of vocabulary distribution will, with time, tend to the limiting form of random partitioning which is subject to chance only. In particular, the ergodic property of the system of vocabulary distribution is such that if the total text is divided into an arbitrary number of equal parts, the number of words peculiar to a part, common to 2,3 ... all parts will be sensibly the same for each such part, or combination of parts, and the alternative probability of the extremes, i.e. the words peculiar to, and words common to all, parts will depend only upon the total frequency distribution and the number of parts.

20.4 A CHARACTERISTIC QUANTITY OF LINGUISTIC STATISTICS—THE REPEAT RATE

A succinct description of statistics is this: its raw material are the crude facts of experience; from them is evolved the finished product, usually a table of chances, that is a table of probabilities. Linguistic statistics is no exception to the rule. It too starts with the crude facts, i.e. language data obtained from textual, or other analyses, and evolves from them probabilities as the finished product.

1. The statistical analysis of literary texts leads in the first instance to a table of global frequencies, and relative frequencies, of particular types of linguistic forms, e.g. letters, phonemes, morphemes, words. If the samples are great, we may regard the relative frequencies as sufficient estimates of the corresponding probabilities, and in this way the first step of the analysis leads to establishing simple probabilities of occurrence.

If the analysis was not one of texts, but of a dictionary, the probabilities are those of particular categories of word structure comprised in the total vocabulary.

To obtain a characterisation of a linguistic text, or a dictionary, by means of reliable relative frequencies of particular items of a category, is a tremendous task, which —if it had to be carried out bodily, so to speak—would tend to make linguistic statistics rather unpopular. It is, however, the essence of statistics that it manages to summarise experience in a simple and lucid manner.

Such a parameter is Yule's Characteristic K, which for great samples equals the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

sum of the squared probabilities of the particular words, Σp_i^2. This is the overall Repeat Rate for words, namely the probability that two words picked at random from the text will turn out to be the same word.

20.5 THE SUM OF LEAST SQUARES OF LANGUAGE STATISTICS

Suppose we were sampling a text by examining segments of the linear sequence of words representing the text. For the smallest sample, $n = 1$, the relative frequency r of a word $r/n = 1$. As the number of words in the segment increases only slightly to, say, $n = 10$, it will often consist of as many different words, and the relative frequencies of occurrence of a word will be $p_1 = p_2 = \ldots p_{10} = .10$. A substantial increase in sample size will have the effect that the differences in p_i for high frequency words on the one hand, and low frequency words on the other, will become more and more pronounced. They will, of course, always add up to unity, $\Sigma p_i = 1$, in accordance with the definition of a relative frequency or probability, and this is, incidentally, why their sum cannot be used for characterising particular type-token relations.

However, the sum of their squares, $\Sigma p_i^2 \neq 1$, and may, therefore, be used as a characteristic figure; in fact, the Characteristic K can be interpreted for great samples as the repeat rate of the vocabulary words in a continuous text, $K = \Sigma p_i^2$.

Let us see how the Characteristic K behaves as the sample size and the vocabulary increases. We list below the size of the segment of text examined in terms of number of words and the corresponding K^* for different samples from (a) Pushkin's novel

TABLE 82

	Sample Size (No. of occurrences)	K*
Pushkin, *The Captain's Daughter*	4703	.00663
	4952	.00673
	9147	.00656
	28591	.00603
Occurrences of 490 selected words of Modern Russian	31934	.0052
	33035	.0050
	64145	.0052
	133655	.0046
Frequency of occurrence of 503 selected words of Modern Russian	11720	.00274
	13202	.00260
	28615	.00279
	58130	.00238
Macaulay, *Essay on Bacon*	2000	27.63 (Av.)
	4000	27.33 (Av.)
	6000	27.23 (Av.)
	8000	27.18 (Av.)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

The Captain's Daughter; (*b* & *c*) for chronologically different segments in composite samples of Russian, and for samples from Macaulay's *Essay on Bacon* (Tables 10, 12b, 12c and 11 respectively, Vol. I).

In all examples we find that K, on the whole, decreases with sample size. This systematic decrease should not be confused with the decrease in variation of K as sample size increases. As Table 11 in Vol. I shows, for the four initial samples of some 2000 occurrences each, K ranges from 25.16 to 28.79, a range of 3.63 units; the mean value for the first group is 27.63. For the second group of samples of some 4000 occurrences each, K ranges from 26.25 to 28.47, all of which values fall within the range of the values for the first group; but the mean 27.33 is slightly lower than that for the first group. For the third group based on samples of 6000 occurrences each, the range is from 26.84 to 28.14 and thus again within the range for the second group, and, a fortiori, within the range for the first group; the mean is 27.3, very slightly lower again than that of the second group. Finally, for the total distribution based on about 8000 occurrences the value of K is 28.17 and thus again lower than the mean for the previous group. The fall in the mean value of K, though slight, is continuous from the first group to the last. Thus there is in addition to the decrease in variation with increase of sample size, which is in accordance with the general validity of the Law of Great Numbers, a small but systematic decrease of K to which Yule did not attach any great importance. However, we shall show that this is an essential feature of the Characteristic which leads to an interesting new interpretation of that constant.

The reason for the decrease is not far to seek. As a text proceeds, new words are added to the vocabulary with the result that the old words are less often repeated than they would be without the increase in vocabulary. This leads to a reduction in their relative frequencies. The effect, however, will be slight, the longer the text, since the number of new words introduced also decreases as text length increases. What is more important for the value of K as the sum of the squared probabilities of dictionary words is that the p_i of the high frequency words (e.g. grammar words) will settle down to sensibly constant values as text length increases, whereas the p_i of medium and low frequency words will steadily decrease. This must result in Σp_i^2 approaching a minimum limiting value with increasing text length.

We can, therefore, speak of K for the total text as the *Sum of Least Squares*,[3] meaning that Σp_i^2 for the whole text will be less than for any sample. It follows that the method of characterising the type-token relation in literary texts by K for great samples can be regarded as a technique of estimation in which the quantities under estimate, the series of p_i, are determined by minimising the sum of their squares, and thus as being of the type of least square methods.

Observe the difference between K as a sum of least squares in linguistic statistics and the variance, σ^2, as that of general statistics. The latter is the sum of squared deviations from the arithmetic mean of the variate; the former is the sum of squared probabilities,

[3] Not in the strictly technical sense of the term.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

that is differences of p_t from zero, and *equals* the arithmetic mean, more precisely, the expectation of p_t (Chpt. XIVA).

The proof that K represents the sum of least squares is as follows. Let us suppose that we had divided the vocabulary of a given text into two parts: low frequency and high frequency words, taking as the dividing point an arbitrary frequency of occurrence. As the text grows, the low frequency words will increase in number faster than the high frequency ones which, in fact, will cease adding new words to their number when the text has become very long. This will have the consequence that the number of smaller p_t values, and with it that of smaller repeat rates, will steadily increase with text length. Since all values of p_t must add up to unity, it follows that the high frequency words, and consequently their repeat rates, will decrease as the sample approaches the total text length.

We have numerical evidence for the correctness of that conclusion in the behaviour of the word probabilities in works of different length by Chaucer (Table 5 Vol. I). The increase of the ratio low frequency/high frequency words with text length may be regarded as the general rule of which the observed increase of the proportion of Romance words with text length in Chaucer forms a special case. Romance words were, for Chaucer, words with lower frequency, and new Romance words should, therefore, be introduced at a faster rate than new Germanic ones. This must result in the increase of the proportion of Romance vocabulary with text length.

However, the relative frequency of occurrence of the Romance element remains sensibly constant, 12.5% approximately, and is thus independent of text length. Since the number of Romance words is in the longer texts a multiple of that in the shorter ones, it follows that the average frequency of occurrence must have dropped considerably. The same applies to the Germanic element, though to a lesser degree, since the number of new words introduced as the text grows is much smaller. The result is that the sum of the squared probabilities of particular words tends towards a minimum as texts grow in length.

20.5.1 Conditioned repeat rates

The probability of gaps of a particular length corresponds to the repeat rate of words *if* they are separated by such gaps, and the average probability of gap length for a word, i.e., regardless of the interval length between repetitions, to the general repeat rate. We may regard the repeat rate of words separated by a particular interval as the differential of the repeat rate regardless of interval length.

The iteration of a vocabulary item in speech or writing may be independent of the preceding word, or it may, to some extent, be linked to the presence of another vocabulary item preceding it. In such a case, its repeat rate is conditioned by that of the word to which it is linked.

Another instance of conditional repeat rate can be found in language translation.

If translational equivalents in the source and target languages are arranged in a bi-variate table according to their word length in both languages, they are found to be

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

significantly correlated with regard to the quantitative characteristic of word length. The numbers in the diagonal from top left to right bottom represent the numbers of translational equivalents with the same word length, which, incidentally, are largely responsible for the positive correlation.

The ratio of the sum of the diagonal cell frequencies to the total of translational equivalents in a standard translation of a given work can serve as a criterion of quality of translation in that respect.

The encounter or coincidence of translational equivalents of the same length in the source and target language may be regarded as the repetition of a word of such length, and the square of the probability as a repeat rate. Denoting the repeat rate of a certain word length among translation equivalents as before by p_t^2, their sum, Σp_t^2 represents the *conditional repeat rate* of word length among translational equivalents from the whole of the sample. The summing of the p_t^2 has to be performed over all the numbers in the diagonal, and the resulting quantity determines, largely, the intensity of correlation.

Another instance of conditional repeat rates is the Markoff chain, expressing the extent of repetitions of linguistic forms. For the simple Markoff chain, the coefficient of dispersion is calculated as a difference between vowel vowel (or consonant consonant) and vowel consonant probability. Since the former is the probability of a vowel (or consonant) repetition with zero interval, and the latter the probability of a vowel being followed by a consonant, or vice versa, we may regard the Markoff Chain coefficient as a measure of the extent to which the repeat rate of vowels (or consonants) is interfered with by simple alternation of vowel and consonants in the linear sequence of letters, that is under the conditions of one event being immediately followed by the other.

20.6 THE BINARY PRINCIPLE OF CODING IN THE LIGHT OF GEOMETRICAL PROBABILITY

The different counts of letters, phonemes, vocabulary items discussed so far, are all of a *qualitative* kind, and, insofar, our results refer to linguistic counts on qualitative variables. If the categories of a given type of linguistic form are distinguished *quantitatively*, e.g. by word length in terms of number of letters or phonemes or syllables, or sentence-length in terms of number of words, the ratio of sample counts can here be of two kinds; that of the word category, say, of a given length to all words counted, or that of the number of smaller units, letters say, to all such units in the sample. The latter gives information about the "consumption" or utilisation of the smaller units (letters) by the word categories of different length (Chpts. II, IX).

Using purely linguistic characteristics, it is evident that forms like phonemes or alphabetic signs cannot be analysed in this way since they represent the ultimate linguistic units of the spoken and written language, respectively. Incidentally—and as a sort of compensation—it is precisely the fact of letters and phonemes being the "lesser" units of language, that their probability can be determined not only from

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

texts (as the probability in "la parole"), but also from the dictionary (as probabilities in "la langue"). Similarly, word length may be studied both in literary texts and in the dictionary.

It is now an important consequence of linguistic sampling theory as an application of geometrical probability that it enables us to introduce the "binary" code of information theory in this connection. *From the point of view of geometrical probability the binary code may be said to make the smaller linguistic units, letters and phonemes, capable of being treated like the larger ones as quantitative variables*, with the consequence that the proportion of a particular phoneme, say, in the text, can be expressed by that of a smaller unit used in the manifestation of the phoneme in question, to all such units used in the text.

That smaller unit is the "bit" of the binary code in terms of which any linguistic form can be expressed. This has not only the advantage of providing a common denominator for different *Ty*-systems, thus enabling valid comparisons between them, but it makes the qualitative variables, such as phonemes, quantitative, and thus makes it possible to assess the functional burdening of a phoneme in "bits", or rather the extent to which the "bits" are used by a particular phoneme or group of phonemes. Geometrically, as indicated above, the expression of phonemes, say, in terms of "bits" amounts to a sub-division of the geometrical space they occupy into smaller units, whose number per phoneme is an inverse function of its probability of occurrence. The "bits" are, however, primarily not of a geometrical, but a physical, nature, since their prototype is the number of closings and openings of an electric circuit representing the symbol length for purposes of electrical communication.

This shows the "bits" as not really and directly admitting of a linguistic interpretation. Only indirectly by equating them with certain hypothetical linguistic characteristics can they be interpreted as a property of linguistic systems, except in the case of word length where the "bits" are seen to correspond to the units in which word length is expressed. Otherwise, the linguistic interpretation can only be given via the probability of occurrence and "information" as defined in information theory.

The squared probability of linguistic forms as the differential of linguistic structure is of a descriptive nature, and so are the integrals of p^2, the Characteristic K, the gap distribution, the Markoff Chain coefficients. The Entropy H and linguistic duality, on the other hand, belong to a class of parameters which might be called "critical" in the sense of implying a critical evaluation of that linguistic code, or at least enabling us to apply such a "critique".

The entropy is calculated as the negative weighted mean or, considering that we are dealing here with population probabilities, as the negative expectation of $\log p$

$$H = -\Sigma p \log_2 p = -E(\log_2 p)$$

Its relation to the linguistic repeat rate in the strict sense of the term is as follows. Since $-H = \Sigma p \log p$, which as the arithmetic mean of the logarithmic distribution of p equals the logarithm of the geometric mean of p, we have, using the inequality

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

between the geometric and arithmetic mean of a distribution, $G \leq A = \Sigma pp = K$, and we come to regard antilog $(-H)$ as the lower limit of the repeat rate K (Sect. 14).

Let us suppose we had calculated the entropy H for the word count in two writers, and that $H_1 > H_2$. The conventional interpretation in accordance with the physical implication of entropy is that the vocabulary distribution of writer I was of greater information value than that of writer II. What does that mean in terms of language? I venture to say that among 100 linguists we would not find 10 who agree about the meaning of the conclusion.

On the other hand, let us express the finding in terms of K. It is a matter of simple arithmetic that if $H_1 > H_2$, Antilog $(-H_1) <$ Antilog $(-H_2)$; according to formula (37) XIVA this means that the lower bound of K_1 is smaller than that of K_2. In other words, we may expect a lower repeat rate for the use of vocabulary in writer I. This is immediately understood, provided the term repeat rate is clear. From our knowledge of the form of word count distributions in general we may supplement this conclusion by adding that the word count distribution of writer I has a flatter gradient of probabilities, and thus less structure, which agrees well with H_1 being great, the maximum, H', being the entropy in the case of the equidistribution and thus in the case of the complete absence of gradient.

20.7 THE STATISTICS OF THE STRUCTURE OF MEANING

It is an historical fact that intuitive probability is insufficient for scientific purposes. A case in question was the need for the Bose-Einstein statistics as the "New Statistics" in quantum theory. What seemed to be the appropriate or "natural" probability distribution which up to then had been accepted without hesitation by physicists, turned out to be not suitable when it came to dealing with certain particles like photons and protons. It will be shown that the same theoretical considerations which made the mathematical models of classical probability theory unsuitable for dealing with problems of quantum physics, lead also in mathematical linguistics to abandoning these models and introducing statistics of the Bose-Einstein type. We shall first discuss briefly the mathematical model required, then give the physical background

TABLE 83

1	*abc*	—	—	10	*a*	*bc*	—	19	—	*a*	*bc*
2	—	*abc*	—	11	*b*	*a c*	—	20	—	*b*	*a c*
3	—	—	*abc*	12	*c*	*ab*	—	21	—	*c*	*ab*
4	*ab*	*c*	—	13	*a*	—	*bc*	22	*a*	*b*	*c*
5	*a c*	*b*	—	14	*b*	—	*a c*	23	*a*	*c*	*b*
6	*bc*	*a*	—	15	*c*	—	*ab*	24	*b*	*a*	*c*
7	*ab*	—	*c*	16	—	*ab*	*c*	25	*b*	*c*	*a*
8	*a c*	—	*b*	17	—	*a c*	*b*	26	*c*	*a*	*b*
9	*bc*	—	*a*	18	—	*bc*	*a*	27	*c*	*b*	*a*

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

for the new type of statistics and, lastly, show that the random partitioning function which we use in vocabulary statistics is of the type of Bose-Einstein statistics.

1. The conceptual model for the statistics required is that of the distribution of r balls in n cells. Table 83 describes all possible outcomes of the experiment of placing 3 balls into 3 cells according to the classical conception of probability. Each of these arrangements represents a simple event, or, as we say, a point in sample space which corresponds to the phase space of statistical physics. The total number of arrangements is n^r and, thus, for our example $3^3 = 27$. The probability of a particular arrangement is n^{-r}.

So far, we have assumed that the balls are distinguishable as a, b and c. If we suppose that the balls are indistinguishable, it means that we no longer distinguish between three arrangements such as 4, 5, 6, and Table 83 reduces to

TABLE 84

1	***	—	—	6	*	**	—
2	—	***	—	7	*	—	**
3	—	—	***	8	—	**	*
4	**	*	—	9	—	*	**
5	**	—	*	10	*	*	*

In the scheme above, we have considered indistinguishable balls, but Table 84 stil refers to a first, second, third cell. That is, it still refers to an ordered arrangement of the balls. We can go a step further and assume that even cells are indistinguishable. With both balls and cells indistinguishable, only three different arrangements are possible, namely:

*** — — , ** * — , * * *

with probabilities 3/10, 6/10, 1/10 respectively, each possible arrangement being equally probable, and this is, in effect, the step which leads from classical statistics to quantum statistics.

The mathematical model for the Bose-Einstein statistics assumes both balls and cells to be indistinguishable. This has the consequence that whereas the 10 points of Table 84 have, according to classical physics, the following probability distribution

Point number:	1	2	3	4	5	6	7	8	9	10
Probability:	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{27}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{2}{9}$,

the Bose-Einstein model attributes a probability of 1/10 to each of the 10 sample points.

Generally, the number of distinguishable arrangements if both balls and cells are indistinguishable is no longer n^r, but

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\binom{n+r-1}{r} = \binom{n+r-1}{n-1}$$

all arrangements being equally probable. For our example the number of distinguishable arrangements is

$$\frac{(3+3-1)!}{3!\,2!} = 10$$

each arrangement having the probability 1/10.

2. Attempts to explain Planck's formula of cavity radiation by using Maxwell-Boltzmann statistics proved a failure. Although no pure reasoning could tell why photons would not obey the probability laws which had been found so successful up to then in statistical mechanics, yet it became clear that if the empirical findings were to be established on a statistical foundation, the classical statistics must be completely altered and a new "quantum statistics" set up. The way in which this may be done was shown by Bose in 1924 and improved by Einstein. They assumed complete indistinguishability of the light quanta. We proceed now to work out this type of statistics.[4] For this purpose we denote the individual cells on a given sheet, or level, by $z_1, z_2 \ldots z_{g_s}$; the number of them is, by definition, given by the weight factor g_s of this sheet. On the other hand, let there be n_s particles in the sheet which, for the present, we denote individually by $a_1, a_2 \ldots a_{n_s}$. We have to distribute these particles among the g_s cells of the sheet and determine the number of distinguishable arrangements, keeping both particles and cells indistinguishable. To this end we describe a definite arrangement in the following way. We write down purely formally the elements z and a in an arbitrary order

$$z_1 a_1 a_2 z_2 a_3 z_3 a_4 a_5 a_6 z_4 z_5 a_7 \ldots\ldots$$

with the understanding that the particles standing between two z's are in each case supposed to be in the cell which stands to their left in the sequence. The above sequence means, therefore, that the particles a_1 and a_2 are in the cell z_1; the particle a_3 in the cell z_2; the particles a_4, a_5, a_6 in the cell z_3, no particle in z_4, and so on; that being so, the first letter in the symbolic arrangement must obviously be a z. We therefore obtain all possible arrangements, by first setting down a z at the head of the sequence—which can be done in g_s different ways—and then writing down the remaining $g_s - 1 + n_s$ letters in arbitrary order one after the other. The total number of these arrangements is therefore

$$g_s\,(g_s + n_s - 1)!$$

Distributions which can be derived from one another by mere permutation of the cells among themselves or of the particles among themselves do not, however, represent different states, but one and the same state; the number of these permutations is

[4] M. Born, *Atomic Physics*, 3rd. ed. (London-Glasgow, 1944).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$g_s!\, n_s!$. We thus obtain for the number of distinguishable arrangements in the level, or sheet, which is characterised by the index s, in the case of the Bose-Einstein statistics

$$\frac{g_s(g_s + n_s - 1)!}{g_s!\, n_s!} = \frac{(g_s + n_s - 1)!}{(g_s - 1)!\, n_s!} = \binom{g_s + n_s - 1}{n_s}.$$

Altogether, the number of distinguishable arrangements for the case when there are n_1 particles in the first sheet, n_2 particles in the second and so on, is given by the product of the expressions of the above type for the sheets

$$W = \prod_s \frac{(g_s + n_s - 1)!}{(g_s - 1)!\, n_s!}$$

We call this the "probability" of that distribution of the particles among the various sheets which is defined by the numbers $n_1, n_2 \ldots$. It takes the place here of the probability found in the Boltzmann statistics, viz.

$$W = \frac{n!}{n_1!\, n_2! \ldots} \quad g_1^{n_1}\, g_2^{n_2} \ldots \text{ (See Vol. I, Sect. 18.4)}$$

Without going into further detail, it shall only be mentioned that what is known as the Fermi-Dirac statistics for the distribution of electrons, neutrons and protons can be derived by a further alteration of the assumptions which have led to the model of the Bose-Einstein statistics.

3. Let us now consider the mathematics of the random partitioning function (Sect. 3.3 and 18.5), in the light of the assumptions implied in Bose-Einstein statistics. Let the different vocabulary items stand for "balls" and the samples or parts of the text, or the texts belonging to a group, for "cells". When performing the random partitioning of vocabulary items—all having the same particular frequency of occurrence—, the resulting distribution among the parts of a literary text applies irrespective of the particular vocabulary items which are, therefore, regarded as being indistinguishable. For instance, if we ask how many words of a given frequency group are peculiar to one sample, shared by 2,3, ... r samples, the answer applies irrespective of the particular vocabulary items or irrespective of the meaning of the words, each word belonging to the given frequency group having the same probability of being peculiar to one sample, shared by 2,3, ... r samples.

Moreover, the process of random partitioning by its very definition, treats the different samples or parts of a text as indistinguishable. For instance, if we enquire about the occupancy of the different parts of a text by words belonging to a specified frequency group, the answer required, and obtained, is irrespective of which part or sample we have in mind. It follows that the two conditions which are characteristic for Bose-Einstein statistics, namely complete indistinguishability of vocabulary items as well as of samples or text parts, are essential for the random partitioning of vocabulary. The additional feature, viz. that the vocabulary items corresponding to the particles of quantum statistics, may belong to different frequency groups, also

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

forms part of the Bose-Einstein statistics if they are applied to particles of different energy.

It thus appears that the very same alterations have to be made in the models of classical statistics in order to make them suitable for dealing with light quanta, and with word content as quanta of meaning. It must, however, be emphasised that the use of the random partitioning function in mathematical linguistics is only a beginning of using Bose-Einstein statistics in this field. The consequence of the full application of the "New Statistics" to language problems may be just as far-reaching for "content" as information theory was for "expression". Already at this stage, it has added to the two specifically linguistic parameters of literary statistics, viz., the Characteristic K and the Entropy H, a third: the Vocabulary Partition Constant V.

20.8 LOGARITHMIC PROBABILITIES AND LOGARITHMIC VARIABLES

A characteristic feature of the statistics of language structure is the *logarithmic transformation* of the probabilities.

This should not be confused with the use of the logarithm of linguistic forms, e.g. of text length in its relation to the proportion of specified parts of the vocabulary (Chpt. II) and of word length in terms of number of syllables, letters, phonemes (Chpt. IX). These are transformations of the *variable*, and of the same type as often met with in psychological statistics. Considering that the use of words is a psychological fact, this is what we might expect. It is different with the use of *logarithmic probabilities*, which is peculiar to the statistics of language structure.

Information is defined as the negative logarithm of a probability, and the entropy $H = -\Sigma p_i \log p_i$ as the negative expectation of p_i in the series of p_i. The probabilities p_i may be either those of the individual symbols or of groups of symbols, with the same probability of occurrence.

So far the particular distribution of $\log p_i$ does not enter, except that for efficient coding the p_i should decrease as symbol length increases. For optimal coding, however, the distribution of p_i must be a particular one. Its gradient should be such as to conform with the requirements for a geometric series. This means that $\log p_i$ should form an arithmetic series.

All this applies to the type-token relation, and thus to linguistic forms only. The introduction of the content-cognition relation leads to the combination of the logarithmic variable with the logarithmic probability. The distributions of animal-, plant-, human disease-species according to the number of individuals subsumed under these concepts follows the law of the *logarithmic series*, of which the bilogarithmic linear relation, e.g. between the number of species according to the number of individuals per species, of the number of genera according to the number of species per genus, of the number of towns according to the number of inhabitants, is only a special case (Herdan, 1957).

Under certain conditions, the logarithmic series goes over into the bi-logarithmic

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

linear relation. Yule's law is a stochastic derivation of the particular genera-species number relation which was observed by Willis.

To exhibit clearly the relation between the distribution of logarithmic variables, of the logarithmic probability, and of both, we write them in twin columns:

The distribution of the number (token) of words according to the logarithm of their length (log type) obeys the normal logarithmic law or the "Law of Facility", which makes discrimination of the variable word length easy (therefore its name).	The distribution of the logarithmic probability (log token) according to symbol length (type) is for *optimal coding* a logarithmic series, which makes discrimination between probabilities easy (and therefore represents optimal coding).

The logarithmic series of species according to the number of individuals, and with it the bi-logarithmic linear distribution of the Willis-Pareto type, is a combination of these duals.

We must clearly distinguish between the following types of distribution in language statistics:

(1) The *Zipf* distribution, by which we understand not the spurious bi-logarithmic linear relation between frequency and "rank", but simply the distribution of word occurrence among dictionary words; this is the dual of (2) Yule's distribution of dictionary words according to frequency of occurrence. (3) The distribution of words, vocabulary or occurrence, according to word length, and preferably according to the logarithm of word length. Its dual is (4) the distribution of frequency of occurrence (transformed into the logarithmic probability) according to symbol quality or symbol quantity (length). The combination of log variable and log probability results in (5) the distribution of concepts (species) according to the number of individuals as a logarithmic series, which is a special case of (6) the distribution of concepts (genera) according to the number of subsumed concepts (species) for which Yule devised a stochastic derivation. A clear conception of the different types of distributions which are used for characterising language structure will prevent such confusion as to speak of the bilogarithmic linear relation as the Pareto-Willis-Zipf law, or such neglect of the intrinsic differences between the Zipf "law", Yule's distribution of vocabulary according to frequency of occurrence and the distribution of genera according to the number of species, as occurs in the paper by Simon (1953) who expresses surprise that Yule did not refer to that last named function when discussing word counts.

20.9 PROJECTIVE GEOMETRY AS THE MATHEMATICS OF NON-MEASURABLE PROPERTIES OF LANGUAGE

The considerable admixture of chance in language, when linguistic forms are considered in the mass, makes it possible—and indeed necessary—to use statistical methods

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

for describing language structure. When the linear arrangement of linguistic forms is considered—no matter whether actually "on paper" as the written, or as a sequence of sounds in the spoken language—, another kind of mathematics is used in addition to statistics: one dealing with order and position. It is immaterial whether the investigator, that is the linguist, is conscious of this and whether he expresses his results mathematically; the linguist uses these methods nolens volens as the appropriate ones for structural linguistics.

There are three main methods of the mathematics or order in use by structural linguistics:

(1) The *one-dimensional of the theory of sets*, which is concerned with ascertaining the structure of language as the linear sequence of letters or phonemes (Sect. 7.5). As an illustration we may refer to the analysis of Japanese, conservative and innovating, by B. Bloch. The vocabulary was analysed with a view to ascertaining the order in which the 29 phonemes (28 sounds and 1 "interval") occurred in the language. The result was a complete analysis of phoneme order. The subsequent expression by Harari and Paper in terms of the theory of sets is different from Bloch's analysis only through its being quantitative. In spite of giving only qualitative results, Bloch has used virtually the theory of sets, and thus a geometric method.

Let us imagine a world consisting of one dimension only, that is of a line. A segment of that line which may fitly be called a 'linelander', cannot see anything else but the linelander in front and the linelander behind.[5] All that exists for him are those segments of the line which immediately follow him or which he immediately follows; no linelander can, therefore, ever pass another. Once neighbours, always neighbours. This means that the only possible relations are those of immediate neighbourhood, and this is what the theory of sets puts into mathematical terms.

(2) The *two-dimensional of plane geometry* proper, which is concerned with words as segments of the linear sequence of phonemes, and the system of phonological oppositions, or the distinctive use of phonemes in those segments. This is in essence the work of the distributionalists. One of the prominent features of plane geometry is the point-line duality, and the fact that the principles of plane geometry are used in ascertaining the phonemic structure of a language accounts for the emergence of the type-token language structure.

The world corresponding to a two-dimensional geometry consists of all the configurations which are possible in a plane; a suitable name for the creatures of the two-dimensional world would be 'flatlanders'. Shapes like lines, triangles, squares, pentagons, hexagons and other figures, may move freely about, on or in the surface. If we imagine the flatlanders observing one another, it is evident that they can never see anything but straight lines, because they can never rise above, or sink below, the plane surface, which would be necessary for distinguishing the shapes of their countrymen.

[5] Sect. 17.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

This is precisely how the uncompromising distributionalist considers language structure; no matter what the content is which a word denotes, he sees it only as a particular arrangement of unit elements, phonemes. It follows that the relations between words are also restricted to comparison of their phonemic structure. If they have the same phoneme in common, this corresponds to two lines in flatland crossing one another; two phonemes belonging to the same word are points on one particular line. The total of the relations of words in this sense constitutes the phonemic structure of the language. Content of words need not enter at this stage, and does not enter for the uncompromising distributionalist.

(3) The *three-dimensional of the geometry in space*, which takes content into account, as the third dimension of language structure.

Since the flatlander is, by supposition, capable of distinguishing lines from one another—in terms of language, words of the same and of different structure—he could group the lines according to categories of phonemic structure—and the word occurrences according to dictionary words. This means that the word count distribution could still be regarded as possible in the flatland of the distributionalist.

To realise that a line in flatland corresponds to different shapes, and thus to become aware of these shapes as such, the flatlander must cease to live in two dimensions only and must rise into space, thus becoming a spacelander. In terms of linguistics this means that the content of words is now taken into account, in addition to their being types and tokens. The controversy between the distributionalists and the phonologists of the Prague school is thus seen to be a difference in the dimensionality of geometry; the former admitting only methods of two-dimensional geometry, whereas the latter advocate the use of the methods of three-dimensional geometry.

Apart from phonology, content is considered for its own sake in semantics, and the geometry to apply is that of three-dimensional space. In terms of linguistics, the type-token-content relation acquires here primary importance.

Languages differ with regard to efficiency of coding, and still more as regards the extent to which they avail themselves of the device of linguistic duality. Linguistic laws are man-made, and consequently can be un-made or acted against, just like laws in the legal sense, and with the same disastrous consequences. They are the expression of solidarity of linguistic forms, just like the civil and criminal law express the solidarity of the members of a social community. The extent to which linguistic duality is present in a language or is used in a literary work is a criterion of the "sense of duality" in the language or in the writer.

It should be well understood that the partition function of Sect. 3.3 is also one of geometrical probability. It could, of course, be applied to non-geometrical situations, but it is specially suitable for material which, like linguistic forms, is naturally arranged in geometrical configurations.

In a way, it can be contrasted with binary coding. Whereas the binary code represents a finer sub-division of the geometrical area in the pages of literary text than could be achieved if we were to stop short at the sub-division in terms of the ultimate

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

linguistic units, such as letters and phonemes (Sect. 20.5), the method of random partitioning of vocabulary makes us go in the other direction. Instead of regarding the individual words (vocabulary items) as the uppermost categories under which to arrange the occurrence of linguistic forms in the text, groups of words are now distinguished according to the number of parts of a text to which they exclusively belong.

Again, as before, the partitioning of vocabulary as a matter of pure geometrical probability, will suit the actual distribution of vocabulary only if, and insofar as, that distribution is random, that is, governed by chance.

Although we speak of three types of geometry which are used in structural linguistics, it is not the whole of either which is required in linguistics, but only certain selected features. Let us now consider which these features are, and whether they have something in common that would make them specially suitable for the study of language properties.

The geometry of sets as we apply it in linguistics deals only with ordered pairs of linguistic forms, immediate neighbours. Plane geometry when applied to linguistics as distributionalism is concerned with the type-token relation, be it of vocabulary or text material; its main research tool is thus the linguistic duality. The same is true for the investigations which take "content" into account, with this difference that the simple type-token duality now gives place to the type-token-content duality.

It is evident that what the three special features of the mathematical structure of language, epitomised by *K*, *H* and Duality, have in common is a dyadic or twin element. We may speak of repeat rate, binary code, and geometrical duality as the *three-fold root of the dyadic structure of language.*[6] There is, however, an important difference between linguistic duality on the one hand, and the descriptive parameters and the entropy on the other. The descriptive parameters are measures in the strict sense of the term, since they belong to the area of statistics, and statistics as today defined is the theory of measurement. The entropy can also be described as belonging to the category of measurement, although the binary units in which it is expressed are an artefact by which we *make* the entropy numerical. Duality, on the other hand, is not a measurement. It is a concept of projective geometry, one of whose essential characteristics is that measurement plays no part in it. This is expressed by the German term for it: "Geometrie der Lage".

[6] If symbolic logic is considered as a part of mathematical linguistics, Boole's algebraic "law of duality" would make us speak of the "four-fold root of the dyadic structure of language" (The Epilogue).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

20.10 SPECIFICITY OF PARAMETERS AND METHODS

One source of confusion in mathematical linguistics is the mathematician's desire for generalisation. Admittedly this is one of the beauties of mathematics, but in linguistics it is sometimes out of place.

An instance of such out-of-place generalisation is Simon's stochastic model for what he calls Yule's law and the Zipf law. What Simon calls Yule's law is Yule's formula for the relation between the numbers of genera and species. He expresses surprise that Yule did not think of his own work when writing his book on the statistical study of vocabulary, but the explanation is very simple: the word frequency distribution for which Yule in his book devised the Characteristic K as a suitable constant is very different indeed from what Simon calls "Yule's law", which, by the way, is again different from Williams' distribution of species according to the number of individuals. Again, Yule's word count distribution is by no means identical with Zipf's distribution where frequency of occurrence is plotted against "rank". Simon treats all three distributions as if they were more or less the same. Hence illae lacrimae (as shed in Sect. 15.3).

The need for discrimination in mathematical treatment applies in a still higher degree when different levels of expression are concerned. It is not denied that the relation between all these levels entails relations between the mathematical characteristics or between the statistical parameters designed for the various levels, but the indiscriminate extension of such parameters designed for one particular plane of expression to any and all such planes is, as a rule, not satisfactory and often justifies criticism by the linguist.

Thus, v_m as a stable value, characterising the type-token relation irrespective of sample size, can also be used if "type" stands for phonemes; but since the number of phonemes is constant for a given language in samples of any size, N in formula (28) Chpt. V is constant, and not only v_m, but the cofficient of variation v_x is constant. There is, therefore, strictly speaking, no particular need for working with v_m. Such need arises when "type" stands for the words of a language; since the vocabulary, N, changes from sample to sample, and especially with sample size, no statistic except v_m is a constant of the distribution.

The logarithmic distribution of linguistic forms is specially suitable for word length, in terms of syllable, phoneme or letter number, also for sentence length in terms of word or syllable number, and for occurrence frequency of vocabulary items. On the other hand, the logarithmic transformation of the probability is used in information theory.

The distribution of gaps between repetitions of given words in a text may be used for dictionary and grammar words alike, but the method will be found more valuable with the latter; the repetitiveness of the former is better characterised by the repeat rate. and by the pattern of vocabulary connectivity.

Instead of hasty generalisations, it is advisable to study the relation between the

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

parameters devised for different planes of language, because such relations may reveal something linguistically relevant about the different levels. As an example, we mention the relation between H and K (Chapt. XIVA).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

XXI

THE "EIGENSTATISTICS" OF LANGUAGE

21.1 EPISTEMOLOGICAL NATURE OF THE PRINCIPLE OF LINGUISTIC DUALITY

The mathematical characteristics of linguistic events all have the dual element in one form or another: as repeat rate, as binary code, as geometrical duality. We shall suggest that this dual element epitomises a fundamental law of symbolic expression which, with regard to linguistic content, was formulated by G. Boole as his algebraic "law of duality", and which, with regard to linguistic expression, we have described as the geometrical law of duality in language. That we cannot as yet strictly prove the fundamental identity of these laws and of the other form of duality in language, does not make a theory of this type "mystical" as one mathematician (Oettinger, 1957) seems inclined to believe.

First, what he means cannot be "mystical" but perhaps "metaphysical". "Mysticism" according to the Oxford dictionary is "the philosophical teaching that the Deity or the Absolute can be perceived by intuition", which has evidently nothing to do with the explanation of certain language properties by reference to an all-pervading duality. "Metaphysics", on the other hand, is defined by the Oxford dictionary as "a philosophy aiming at an integration of the outer and inner experience so as to arrive at an harmonic picture of the world, which does not exclude the reference to a transcendent entity, i.e. beyond our experience". But even "metaphysical" would be a misnomer in the given case because our theory of the dual element in language consists in tracing a law of symbolic thought in the picture of the manifold in terms of language, and is thus decidedly *epistemological*, and as such, non-metaphysical. As a condition for conceptual thought in linguistic expression, duality occupies a position analogous to that which space and time occupy, according to Kant, in the world of outer experience.

But more: not only is the conception of an all-pervading principle of duality in language neither mystical nor metaphysical in the sense of being non-scientific but, considering that the various forms of duality in language as established by Boole (law of duality), Trubetzkoy (phonological opposition), Jakobson (distinctive feature opposition), Information Theory (binary coding), Language Statistics (repeat rate) and by myself as Linguistic Duality were all established independently, it would be rather unscientific to leave it at that and not aim at a unification of these forms by reference to some common cause. Our contention is that there exists a relation between the twin-principle of language and the uni-laterality of the speech function in

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the brain. Most likely, it is the fact of duality being intrinsic to language which makes symmetry of the speech function in the brain redundant.

21.2 STATISTICS FROM "WITHOUT" AND "WITHIN"

An important feature of the calculus of linguistic observations is that certain concepts of structural linguistics should prove to be of a mathematical nature, by which is meant that they correspond to certain mathematical concepts. This enables us to transfer particular sets of mathematical axioms directly to language data. Thus, the interpretation of 'la langue' plus probability structure of word occurrence as the statistical population and 'la parole' as samples from it, is the foundation for applying statistical methods to language.

However, it is one thing to conceive of certain concepts of linguistics as being of an essentially statistical nature and, on the strength of this, apply statistical methods, and another to maintain that the very structure and practice of language—quite apart from language as an object of scientific study—implies the unintentional use of mathematical methods.

In the first instance, mathematics are intentionally superimposed upon linguistic material; in the second they are unintentionally used because they are built into, and thus form part of, language.

This is the function of linguistic duality; another instance is the distinction between phonemes and variants so essential for language, which corresponds to that of significant and non significant differences of the theory of statistical inference. Still more striking are the 'Eigenstatistics' of language on the vocabulary level. Whereas on the phonemic level it is only the basic ideas of statistical inference that are used, we find on the vocabulary and content level a rather complicated design for purposes of comparison of the Variance Analysis-type. This is one of the truly remarkable instances of using a complex quantitative design, involving constant parameters, without the speaker being aware of it. As such it is reminiscent of the *HP*-content of the blood of which we are quite unaware, but whose maintenance is essential for our well-being. Our organisation, therefore, strives to maintain it within narrow limits of variation without our being in the slightest conscious of it.

I apologise to linguists for explaining in the following paragraphs something which is quite elementary. However, readers other than linguists may find it welcome.

21.3 THE PHONOLOGICAL OPPOSITION

Language according to Phonology consists of a number of elements whose essential characteristic is that they differ significantly from one another. A word must differ from all other words by something, since otherwise it would not be recognised as that particular word. There exists only a limited number of such elements for distinction, and since that number is much smaller than that of the words to be formed out of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

them, the words must consist of combinations of those elements. But not every possible combination is admissible; the combinations are subject to special rules, which differ from language to language.

What we have called "elements" are sounds, but not every sound difference is connected with a difference in meaning, and it is the object of Phonology to find out which differences between sounds have what we call a *distinctive function*, and which are merely *speech variants*. The former are called *Phonemes*, the latter *phonetic variants*.

The concept of differentiation between objects presupposes that of comparison or opposition. Differentiation in nature is duplicated—though by no means paralleled —by differentiation of word structure. A sound has distinctive, i.e. differentiating function, only insofar as it is in opposition to another sound. It follows that phonemes as sounds with distinctive function are always members of an opposition, and such oppositions are called *phonological* or *distinctive oppositions*. The phoneme inventory of a language is conditioned by the system of its phonological oppositions. It it not the phoneme which is primary, but the distinctive opposition. A phoneme has distinctive function only insofar as it is a member of a phonological opposition.

It is by means of such oppositions that we distinguish words as having different meaning. We shall not go into details about how to determine the distinctive oppositions, and thus the phonemes in a language, which enables us to separate the phonemes from mere variants, but shall mention only the basic rules. If two sounds of a language occurring in the same sound neighbourhood may change places without changing hereby the meaning of the words, then the sounds are only phonetic variants of the same phoneme. The uvolar and the tongue *r* are certainly two different sounds, but the difference is one of habit or taste and is not used for the purpose of semantic differentiation. As another instance where sound difference is not tantamount to semantic difference, may be mentioned *l* and *r* in Chinese. These sounds, quite distinct to our ear and most important for semantic differentiation, are interchangeable in Chinese pronunciation which means that they have there no semantic import.

If, on the other hand, the sounds occur in the same neighbourhood and cannot change places without changing the meaning of the words, or making the words unrecognisable, then they are realisations of different phonemes.

21.4 THE "EIGENFUNCTION" OF LINGUISTIC CONTROL—LANGUAGE AS HAVING COMPENSATORY DEVICES FOR THE ONE-SIDEDNESS OF THE SPEECH FUNCTION IN THE BRAIN

The element of opposition being thus inherent in the structure of language, or rather distinctive opposition being the very soul of language, there is no need for opposition or differentiation to be introduced from without. Language is "self-supporting" in this respect, and does not require two-sidedness of the brain function to be imposed upon it. Indeed, it is obvious that two-sidedness of the speech function would lead to nothing but chaos. The phonological distinctions are too specific to be provided by

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

a mechanical two-sided activity of the brain. Paraphrasing the saying about having a dog and doing the barking oneself, it is a case of "why have a dog if in any case one has to do the barking oneself, and does it better!"

The self-sufficiency of language in providing its own system of controls of opposition is in contrast to the need for bi-laterality of the motor and sensory activity of the brain, but comparable to its intellectual activity in providing artificial patterns for comparison. Empirical investigation is in general based upon comparison. This is especially true for experimentation. If we wish to ascertain the effect of a new drug on Guinea pigs, we compare the treated animal with an untreated one, as control. Since biological variations between animals may interfere with the effect of the drug, we must consider the variations of the effect in treated animals in the light of what may happen in the controls, before regarding the results as significant effects of the treatment. As is well known, both the design of experiments and the evaluation of results by statistical mathematics have, in recent years, been raised to the level of a fine art. What interests us here is only the fact that comparison of experimental and control animals (or plants or whatever the object of investigation) is the soul of statistical inference. Apparent differences are examined with a view to determining whether they are significant. Mathematically this means that, in general, we start with more or less different facts, and finish up with a table of probabilities from which to decide whether the observed differences are significant.

Much ingenuity is required to design experiments so as to have controls just where and how they are wanted. For instance, if a new eye treatment is to be tried on the rabbit, the ideal control for a treated eye would be the other eye of the same animal, because this does away with the factor of biological variations between different animals. If, however, two or more treatments are to be tried in the same set of experiments, or on the same group of rabbits, we are up against the fact that the rabbit has only two eyes. It was now up to the mathematician to overcome the discourtesy on the part of nature in not providing the rabbit with as many eyes as the experimenter would wish him to possess by devising a suitable design of experiment—and he has responded splendidly, by devising what is known as the Incomplete Balanced Block design.

In language, on the other hand, all this and much more in the way of combinatorial technique, designed so as to make language work as it should, is already embodied in the system of phonological oppositions. Shannon (*Printed English*, 1955) remarks about the amount of statistics used in speech, but he means by it only frequencies (probabilities and transition probabilities), whereas I maintain that the speaker uses also a complete significance testing apparatus which is implied in language structure. The versatility of distinctive oppositions may be inferred from the elaborate system of classification of oppositions sketched in Sect. 7.6.

The nature of language as a system of oppositions means that it is the "significant differences"—not in the statistical, but in the semantic sense—which are given. By "given" we mean that it is they which are required first for learning a language through

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

lexicon and grammar. The "sample occurrences" of these oppositions, that is their phonetical realisation in speech, come after we have learnt to distinguish between the members of the distinctive oppositions. Even if psychologists should take exception to this statement as far as the order of events is concerned, yet it remains definitely true for language as such: its primary or basic feature is the system of phonological oppositions, and its secondary or accidental aspect, the phonetic realisation of that system in actual speech. The system of phonological oppositions is analogous to that of statistical controls, but the linguistic sequence: phonological opposition → realisation in speech is the reverse of the sequence of statistical inference: sample value → population value.

This shows how justified was the question put some time ago by Professor Hoenigswald, "How unstatistical is linguistics?" Our answer is very unexpected. *Language starts with what statistical inference ends with*, namely, the significant differences themselves—though not in the statistical, but in what G. Boole would call the logical sense.

21.5 THE STRUCTURE OF MEANING—THE INCOMPLETE BALANCED BLOCK DESIGN OF VOCABULARY STRUCTURE

21.5.1 The system of controls in experimental design according to the analysis of variance

The classical method of experimentation aimed at physical separation or isolation of the variables in which the investigator was interested. Such isolation of variables is, to a great extent, possible if the work is carried out in the physical or chemical laboratory. However, it would be in vain to attempt isolation of this kind in biological or psychological experiments, where the events are, as a rule, due to a complex of causes. The realisation of this gave rise in recent times to principles of experimentation which may be regarded as the opposite of the classical way. The modern way of experimental design not only does no longer aim at the strictest possible isolation of variables, but actually encourages variation in results through including all possible concurrent causes. The justification for this method lies in the fact that we have learned how to separate mathematically that part of the effect which is due to particular causes. This, however, requires careful design of the experiment.

The principles of such experimental design may be briefly stated as follows. In order to assess the experimental error, we require a number of experiments carried out under the same conditions. The allocation of experimental units to the different experimental conditions which we would like to compare as regards the effect must be random so as to exclude bias. In order to provide as broad as possible a basis for our conclusion; and to increase the sensitivity of the experiment (reduce the experimental error) "blocks" of experimental units are formed. This means that all the different experimental conditions which we wish to include in the design are to be tried on groups of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

experimental units, which must be as much as possible similar to one another within a block.

Let us now recall the results of Sect. 3.2.

For the three series of samples, we found these uniformities:

1. For any sample *A*, *B*, *C*, *D*, the ratio of the sum of words common to all four samples in the series and words peculiar to that sample to all the words in the sample was something between 55% and 60%.
2. The averages of the number of words common to any two samples, e.g. $(AB\gamma\delta)$, and of the number of words common to any three samples, e.g. $(ABC\delta)$, were considerably smaller than either of the two other groups $(ABCD)$ and $(A\beta\gamma\delta)$; moreover $(AB\gamma\delta) > (ABC\delta)$.

We now express these results in a slightly different way. To the extent of the number $(ABCD)$, the samples are alike, and thus conform with the idea of the randomised block design: of the $(ABCD)$ words, each is given a "trial" in each sample.

With regard to the $(ABC\delta)$ and $(AB\gamma\delta)$ numbers, the samples are partly alike —insofar as a word occurs in three samples only or in two samples only, but in such a way as to make the number of such words ideally the same for any two or three samples respectively, which is the idea of the incomplete balanced block design (p. 317).

With regard to the $(A\beta\gamma\delta)$ numbers, on the other hand, the samples are completely different. They do not seem to conform to any feature of the Variance Analysis design. But we have yet to consider the role played by the constant probability in a sample of words common to all samples plus those peculiar to one sample. It is now the most remarkable fact, and as such lends support to the statement we made in the heading of Section 21.4, that the probability in question has a controlling effect such as to fit into the total design: as the $(A\beta\gamma\delta)$ number increases, so does the $(ABCD)$ number, and also the two other classes $(ABC\delta)$ and $(AB\gamma\delta)$, all in such a way as to keep the proportion in question constant.

21.5.2 *The distribution of vocabulary in literary samples in the light of the principles of experimental design*

The principles of linguistic vocabulary design may now be formulated as follows.

There are words occurring in all samples from a given total, be it one text or a number of texts by the same writer. If that were to apply to all words, it would produce a great monotony of style and an all-too-obvious identity of content in different parts of the text: the samples would be more or less replicas of one another and of the whole.

Instead, therefore, of keeping strictly to the idea of letting every word appear in all samples, the vocabulary is partitioned in a judicious or balanced way so as to let pairs of samples have an exclusive part of the whole vocabulary. Balance is achieved by letting each such pair have approximately the same number of words peculiar to it.

The same device for achieving a controlled diversity of vocabulary in different parts

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of a whole literary unit is applied to the triple sample groups, the number of words common to any such group being slightly but systematically smaller than that common to the two-sample groups.

A difficulty arises over the words peculiar to one sample. Such words are absolutely necessary for true diversity of style and variation of content between the parts; but they have no counterpart in the designs of controlled experimentation, since uniqueness is against the principle of repeated trials.

However, language has apparently solved the problem of how to include single observations as part of controlled experiments. The vocabulary structure is such that the number of one-sample words plus that of all-sample words is a constant proportion of all the words in the sample, the proportion differing according to the number of parts in which the whole is divided. As the one-sample words become more numerous, so do—almost at the same rate or a little less—the all-sample words, and the two- and three-sample words, the increase throughout being such as to make the proportion All- plus One-sample words/All words in the sample a constant for the literary unit.

This suggests that whereas in other fields such as economics or agriculture, the experimental design needed for the analysis of variance is imposed upon the crude data from outside, so to speak, in language statistics an order comparable to the analysis of variance design is the result not of conscious or 'goal-directed' application of Fisher's method, but forms an intrinsic part of the very mechanism of language, written or spoken, itself, and thus acts from inside.

21.5.3 The structure of meaning as a statistical invariant

For the understanding of the term "structured meaning", it is essential that we should realise that whatever a given text is about, it represents a linear sequence of linguistic forms, and, if the spaces between words are considered part of the symbolism, a continuous line which, for reasons of convenience, is broken up into smaller ones. This is how the text appears in the *E*-plane.[1] The question now arises whether the corresponding event in the *C*-plane[2] has a structure of its own differing from that of the event in the *E*-plane. If such a structure exists, it can fitly be called a structure of meaning, as different from that of expression.

Under a literary text we understand in this connection the exposition of a given subject, and not just the repetition ad infinitum of one word or one sentence, in which case the term "structured meaning" would be obviously inapplicable. The more the text is devoted to the development of one idea, the more will the temporal sequence of the language symbols required for its expression differ in structure from the plan of the book as it has taken shape in the author's mind. Such a plan is intrinsically timeless, since the whole idea must always be present in the author's mind, and must have been there before he started writing it up, or as we say, bringing it to book. It is this plan which we ought to think of when speaking of the structure of meaning, as the

[1] Plane of expression.
[2] Plane of content.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

architectonic distribution of thought prior to its expression. This is in full agreement with Hjelmslev's concept of the structure of meaning (Hjelmslev, 1958).[3]

It is now our contention that this thought-structure of the *C*-plane is reflected in the *E*-plane by the curious distribution of vocabulary among the parts of the text which we have called the vocabulary connectivity. It follows, moreover, from what has been said in Sects. 21.5.1, 21.5.2, that the structure of meaning as reflected by the pattern of vocabulary connectivity must be very similar to the system of thought underlying modern experimental design.

The principle of evaluation of results from the Analysis of Variance consists in comparing the variances of the different columns, or rows, which, according to the experimental design, might reveal special factors influencing the event under examination. The comparison of the variances is only an auxiliary tool for obtaining information about significant diversity of the averages of parts of the experimental series. Roughly speaking, the experiment is so designed that a more or less plausible, but hypothetical, uniformity among the values of the variable (the null-hypothesis) may have to give way to significant diversity.

In vocabulary statistics of the kind described in Sect. 3.2 on the other hand, we have before us

a) four samples from one particular literary text,
b) four small texts of the same literary genre by the same author,
c) four different works by the same author.

The current view on matters of style would be that there must be diversity between the samples, and increasingly so as we proceed from a) to c). Yet as the statistics of vocabulary distribution show there is a far-reaching uniformity, not only "within" each series, but also "between" them. This we interpret as being brought about by a general method of "text-building", regardless of content, which may be regarded as the linguistic counterpart of the null-hypothesis.

There is the class of words which occur in all samples of a series; it forms the basis or fundament of the structure. Then there are the classes of words belonging to any three, and those belonging to any two samples of the series, which provide an incomplete linkage between a) any three, and b) any two of the samples; they have the function of brackets. Finally, there are the words which are unique to a sample, and thus "without the law" of vocabulary solidarity, were it not for the constant ratio between the sum of the all- and one-sample words to all the words in the sample, which maintains the balance between one-sample words and the rest of the vocabulary and thus secures a modicum of solidarity. To summarise: the balanced incomplete block design as appropriate to language (vocabulary) is such as to bring out *real uniformity underlying apparent diversity.*

The fact of the invariance of the numerical relations, no matter whether the samples within a series are random samples in the technical sense of the term (series *a*) or not

[3] L. Hjelmslev, *Proc. 8th Congress of Linguists, Oslo, 1958*, p. 642-5.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(series *b* & *c*), is of great importance. Although the random partitioning of vocabulary gives a better fit for the vocabulary distribution of series *a*, yet the basic numerical relations *within* a series are the same for all three series. Thus, structure in this sense is perfectly general and independent of the subject. It, therefore, constitutes what linguists have called the structure of meaning, which is thus shown to be of the nature of a statistical invariant.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

EPILOGUE

A reviewer of LANGUAGE AS CHOICE AND CHANCE (R. W. Brown, in Language, Vol. 33, pp. 170-181, 1958) while admitting that geometrical duality may be a characteristic feature of language, asks what is the good of viewing language in terms of duality—a question perhaps not quite unjustified, considering the way in which the concept of duality was presented in Vol. I. Now, however, the answer can be given in unequivocal terms: it is through that concept that mathematical linguistics, as presented here, appears to be an extention of G. Boole's LAWS OF THOUGHT (1854) so as to cover not only language 'content' (as in symbolic logic), but also language 'expression' (as in statistical linguistics) and the relation between 'content' and 'expression', as that between Type and Token (as in Parts IV of Vol. I and this book). Moreover, as will be shown, this does for the comparison of stages of linguistic development what stylostatistics does for the comparison of individuals as regards the use of language, and thus provides a new tool of linguistic research.

In what follows it will be shown that what corresponds in linguistic expression to Boole's fundamental law of duality in symbolic logic (I), is the Repeat Rate of linguistic forms embodying the 'Langue-Parole' or the Type-Token duality, as the best characteristic of linguistic structure (II), and that the relation between different states or levels of language, or better, the different linguistic dimensions, is governed by the dimensional duality known in higher mathematics as the principle of duality (III).

I

G. Boole found that "the laws of the symbols of logic are mathematical in their form, and that they are actually developed in the essential laws of human language" (Boole, 1854, pp. 45, 46). As the fundamental law of thought he formulates what he calls the law of duality

$$x = x^2 \tag{1}$$

from which we have immediately

$$x(1 - x) = 0 \tag{2}$$

which is the algebraic expression of the logical principle of contradiction. For let us

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

give the symbol x the particular interpretation 'men', and let 0 and 1 stand for 'nothing' and 'universe' respectively, then $1 - x$ will represent "everything in the universe which is not 'men', or the class of 'not men'." Now the product of the expression of two classes represents that class of individuals which is common to both. Hence $x(1 - x)$ will represent the class whose members are at once 'men' and 'not men', and equation (2) thus expresses the principle that a class whose members are at the same time 'men' and 'not men' does not exist. In other words, it is impossible for the same individual to be at the same time a man and not a man. Now let the meaning of the symbol x be extended from representing 'men' to that of any class of beings characterised by the possession of any quality whatever; and equation (2) will then express that it is impossible for a being to possess any quality and not to possess it at the same time.

But this is identically what philosophers have called the Principle of Contradiction, which affirms that it is impossible for any being to possess a quality and at the same time not to possess it, and which Aristotle has described as the fundamental axiom of all philosophy.

Boole then arrives at the conclusion that the laws of logic are identical with those of algebra, with the single addition that the symbols of logic are further subject to a special law, $x^2 = x$, to which the symbols of mathematics are not subject. Thus in common algebra it is not true that every x is equal to its square—it is true only for $x = 0$ and $x = 1$—, whereas in the Boolean algebra of logic this is true. Boole believes that if that law of thought had not existed, the whole procedure of the understanding would have been different from what it is.

Thus it is a consequence of the fact that the fundamental equation of thought is one of the second degree that we perform the operation of analysis and classification into pairs of oppositions or, as it is technically said, by dichotomy.

In intimate alliance with that law of thought which is expressed by an equation of the second degree, and which Boole has called the *law of duality*, stands the tendency of ancient, and modern, thought to those forms of philosophical speculation which are known under the name of Dualism. E.g. in the system of Pythagoras which seems to have been a combination of Dualism with other elements derived from the study of numbers, ten fundamental antitheses are recognised. Other examples of the pattern of thought of Dualism are systems embodying the oppositions God and the Devil, good and evil, matter and spirit, body and soul, will and idea. The wide prevalence of the particular theories of Dualism together with their manifest analogy with the expressed laws of thought may be conceived to indicate a connection between the two systems. Operation of that law of thought, the law of duality, may have its own peculiar risk of error exalting mere want of agreement into contrariety, and thus form a world which we necessarily view as parts supplementary to each other, permitting the conception of a world fundamentally divided by opposing powers. Such may have been the origin of Dualism,—independently of the question whether Dualism is in any form a true theory or not. It is simply a fact that we cannot proceed very far

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

in our thoughts without it and that sooner or later we shall arrive at systems of dichotomy.

This shows R. Jakobson's binary theory of phoneme differences in a new light. It would seem to be a consequence of the tendency which leads to Dualism, in one form or another. Whether the duality is also inherent in the phonemic system as a physical system, or whether it is imposed by our own way of thinking, without having a physical counterpart, remains an open question, and one which, so far, no philosophy has succeeded in answering satisfactorily. The fact remains that if we wish to understand the phonemic systems as used in different languages, we cannot but ultimately arrive at a binary theory such as devised by Jakobson.

In accordance with the fundamental importance of the law of duality in logic, and therefore also in language as the expression of conceptual thought, it should be possible to replace any linguistic code by a binary one, which is what has been done in information theory. Rapoport (1959) takes exception to my contention that it is the binary code used in information theory which makes the mathematical model suitable for, and applicable to, language. He believes this is like saying that the basic principle of arithmetic is that all numbers can be expressed in the decimal system. Here is just another instance of the mathematician or electronic engineer not paying enough attention to what is required if his 'models' are to be applied to language in a truly sensible way. He acts on the tacit assumption that if mathematics are applicable to language they must be so lock, stock and barrel, which however is contrary to the facts.

It is, of course, true beyond doubt that the formulae for physical entropy as used in statistical physics could be written in logarithms to any base. But as it happens, information theory works with a binary code, which results in expressing the entropy in terms of logarithms to the base 2, and this for very good reasons. Such a code is needed for, and implied in, the interpretation of the entropy as information, namely as the number of guesses required for obtaining an answer to a question. More precisely, we can say that *the formula for physical entropy if it is to be used as a measure of information, admits of only two values of the variable.* And this is in perfect agreement with the limitation which G. Boole found it necessary to impose upon algebra in its application to logic, viz. that if the rules of algebra for the combination of variable quantities are to be applied to logic, the variable could take only two values, 0 and 1. It is only under this restrictive condition that logic can be shown to be essentially of a mathematical nature. A consequence of this condition is the relation which Boole has called the law of duality and which he regarded as quite fundamental for the application of mathematics to logic.

II

The opposition expressed by Boole's law of duality is what logicians call a contradictory opposition, to distinguish it from the contrary opposition (opposition of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

contrariety) between the extreme concepts of a graduated series of qualities. The contradictory opposition between presence and absence of a quality is purely formal, and may be called the logical opposition, whereas that between the extreme concepts of a series of qualities which form the gradual transition from one extreme to the other may be considered a qualitative and characteristically linguistic opposition. It is true that words cannot always be regarded as extremes of a series of transitional qualities—though it is not very difficult to construe matters so that they appear to be—, but insofar as they express qualities or aspects of the objects in the manifold they stand to one another in the relation of qualitative opposites. The description of parts of the manifold in terms of language is epitomised by the number of different "extremes"—words—in relation to the total number of words, and the most useful parameter of a text is the Repeat Rate, that is the overall rate of repetition of particular vocabulary items in a given text, or universe of discourse.

The often heard objection against this way of summarizing text properties is that a word can be used in different meanings and that, therefore, the number of occurrences ought not to be regarded as something homogeneous. But this objection rests upon a misunderstanding. In using a summarizing parameter such as the Repeat Rate or the Characteristic, we are not concerned with meaning at all, but according to the definition of the parameter only with word occurrence and non-occurrence. Just as in Boolean algebra, the variable can take only two values, 0 and 1, so in our algebra, word-count algebra, we are only concerned with whether a particular word occurs or does not occur.

The fundamental statistical parameter epitomising the relation between uniformity and diversity of expression is the Characteristic K, the ratio of the second moment of the variable to the square of the first moment, which for great samples approximates the Repeat Rate:

$$\underset{N=\infty}{K} = \Sigma p_t^2 = E(p) \tag{3}$$

where p_t is the probability of a word (vocabulary item), Σ the sign for summation and $E(p)$ the expected or mean value of p.

The relation between the 2nd and 3rd term, i.e. between the repeat rate (sum of squared probabilities) of a certain class of linguistic forms, say, words, and the expectation of their probability, is seen to be the statistical equivalent of equation (1): just as the relation (1) embodies a fundamental principle of logic or linguistic content in the form of an equation of the second degree, so does relation (3) embody a fundamental principle of linguistic expression in a completely analogous form. This means that a basic statistical parameter in the field of language is of the same form as the fundamental law of symbolic logic, and considering both, the similarity and the difference between (1) and (3), we may speak of (3) as the *statistical form of the law of duality*. As such, it does for language expression (words) what (1) does for language content (concepts).

We have thus, here again, as in the case of the fundamental law of logic, an equation

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

of the second degree—this time in cumulative form, as a statistical equation—and we may insist that if it had not existed, the whole procedure of expression by language would have been different from what it is. This means that both the logical opposition and the linguistic opposition (the contradictory opposition and the opposition of contrariety) have fundamentally the same mathematical form. It is interesting to note that the mathematical structure of information as conceived in information theory rests upon a combination of the contradictory and contrary opposition.

III

Language is conventionally described as having different strata or levels such as the phonemic, morphemic, syllabic, vocabulary level etc. Clearly, the terms strata and level are only analogies and as such cannot contribute much to our understanding of language as a structure. However, a little consideration will show how these analogies could be transformed or replaced by more adequate and more strictly true concepts with the result that real insight is gained into language structure. All we have to do is to realise that these analogies stand for the mathematical concept of Dimension. Starting from the phoneme as the smallest unit, we have here the points of linguistic concepts, that is the entities of zero dimension. Their combination into unit sequences of any length and definition—morphemes, syllables, words—represent the lines of language, but lines so to speak of the first order. For these linear constructs of language can again be viewed as units of 'la langue', that is roughly as Lexicon items and as such form the geometrical points in contrast with the linear sequence in 'la parole', the spoken or written language. These sequences are then to be regarded as lines of language of the second order, and this continues as the number of linguistic dimensions increases, each 'line' of lower linguistic dimensionality being a 'point' in the next higher dimensionality.

It is of fundamental importance and testifies to the consistency of symbolic logic, statistical linguistics and language geometry, that what the linguistic interpretation of dimensionality as given above amounts to is that for the application of geometry to language the number of dimensions is again reduced to two only: point (zero dimension) and line (one dimension), precisely as in the application of algebra to language the variable can only take two values, 0 and 1.[1]

Having found in I and II duality to be so fundamental a property of language structure, we should expect that here, too, it would come to our aid, i.e. when trying to express the relation between the different language dimensions. Not only is this so, but the duality relation which we encounter here seems more fundamental for language then any encountered so far. Since we are dealing with dimensions, it would seem only natural to look for a dimensional duality, i.e. a duality appropriate to the

[1] The importance of the geometrical form of the 'signifiant' was stressed by de Saussure: "Le signifiant... représente une étendue: cette étendue est mesurable dans une seule dimension, c'est une ligne. ... *Tout le mécanisme de la langue en dépend.*" (*C.L.G.*, 2nd. ed., p. 103, 1948).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

geometrical concept of dimension. Such a duality is the duality of projective geometry known in higher mathematics as the *Principle of Duality*. This is the term for the fundamental property according to which the theorems of projective geometry remain true if 'point' and 'line' are interchanged. This relationship gives projective geometry a character quite distinct from elementary (metric) geometry, in which no such duality exists.

It has now come to light that if we assign to the 'point' of geometry the role of linguistic form (phoneme, morpheme, etc.) which we shall briefly refer to as TYPE, and to the 'line' that of linear sequences of such forms (in words, sentences, texts, etc.) which we shall refer to as TOKEN, there appears to be, on the whole, agreement between the properties of 'types' and 'tokens', as the linguistic 'points' and 'lines', and the formal axioms of geometry. In other words: the statements of geometry remain true in terms of 'type' and 'token' and geometry is transformed into a linguistic theory. The principle of duality thus asserts for language that if in any valid proposition of language the words TYPE (linguistic form, e.g. phoneme, morpheme) and TOKEN (frequency of occurrence, probability) are interchanged, the resulting proposition is also valid. That statement can be extended so as to comprise the relation between TYPE, TOKEN and TEXT, which leads, among other things, to a new conception of the 'structure or content of meaning'. Conversely, the fact of the type-token duality admits the inference that language is a system to which projective geometry applies.

The Type-Token duality enables us to integrate, perhaps for the first time, all the different aspects of language into one overall structure. It applies not only 'between' these aspects, but also 'within' each aspect, which it ought to if it was as fundamental a property of language as we claim it to be. In support of this view may be quoted the dualities between symbol and occurrence distribution for the different linguistic types.[2]

It is advisable to forestall a possible objection similar to that which has recently been raised against statistical linguistics, viz. that statistical uniformities of language and the statistical laws summarizing them are of little linguistic interest, because they simply show language to be subject to such laws as have universal application whenever events are the result of a multiplicity of small causes (Sect. 1.7).

The analogous objection to geometrical duality in language might conceivably be that such relations as have been shown to prevail between Type and Token at the different strata of language are of little linguistic interest because they simply show language as subject to laws having universal application in systems with dimensional structure: whatever the field of knowledge, the relations between different dimensions of a system must always be those of projective geometry.

The answer is again analogous to that given with regard to stylostatistics. Apart from the fact that it is of quite considerable merit in itself to have established language to be such a system, and its state to be subject to the laws connecting different dimensions—a thing which was by no means a priori certain or obvious—, so sweeping

[2] Sect. 17.2.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

a view as that quoted above misses the point that it is precisely the general geometrical nature of language structure which is the prerequisite for establishing significant deviations from it and, thereby, enables us to observe, formulate and evaluate differences between stages of linguistic development, just as stylostatistics enables us to observe, formulate and evaluate differences between individuals in the use of language. As we ascend from the smallest units—letters, phonemes—to larger ones—morphemes, syllables, words, sentences, texts, chronological stages of linguistic development—, the number of linguistic dimensions to be taken into account has the consequence of making the duality relations increasingly complicated. What in the lower ranges of linguistic dimensionality, such as phonemes (points), morphemes (lines), is by and large a mechanical, or automatic, and therefore necessary operation, becomes as we ascend in the number of linguistic dimensions, such as words, sentences, texts, more and more a matter of taste and skill in the use of language, until it reaches the level of an art, in the strict sense of the term. Whereas duality in the lower reaches of language dimensions is common to all users of the language, it becomes less and less so as we ascend in the scale of linguistic dimensionality, until the regions are reached where the proper use of duality is a matter for the few only.

In the higher ranges of linguistic dimensions there is the possibility of a deliberate misuse of the principle of duality as I have shown (Chpt. XVIIIA and Vol. I Sect. 15.3) which makes duality in language a most valuable diagnostic criterion of what may be called the social pathology of language. This may be regarded as the social counterpart to stylostatistics: it does for the use of language in a community what stylostatistics does for the use of language by the individual.

IV

Both laws of duality, the algebraic and the statistical, are properly speaking code laws: they ensure the efficiency of language as a symbolic code for describing the content of the manifold.

The principle of contradiction, or, algebraically, the law of duality (1) is a necessary consequence of concept formation which, in itself, is the means of limiting the number of code signs in the language, by using one sign for all the individually different objects subsumed under the concept. Thus, the law of duality as formulated by Boole, as the necessary consequence of conceptual thought, effectively limits the *number of signs* required in the linguistic code.

The statistical law of duality (2) carries the process of limitation a step further by establishing a differential frequency of use of words in speech. It thus embodies the limiting factor in the *use of code signs*.

Just as in the conceptual description of the universe the quadratic term in (1) is operative, being the algebraic expression of the identity of objects subsumed under the concept for the purpose of symbolic thought, and so does away with the necessity of a different sign for every object, so it is in the use of vocabulary the quadratic term in

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

(2) which as the repeat rate of particular words reveals itself as the limiting factor, through doing away with the necessity of increasing the number of different words (and concepts) indefinitely.

So far the relation between the manifold of experience and its expression in language has not been considered, though it is clear beyond doubt that this is a most essential matter for any linguistic code. Since the expression of the content of our consciousness can only be in terms of code symbols, and since the relations of these symbols are governed by the laws of thought—which are not identical with the laws of nature—, a certain amount of discrepancy between content and expression is unavoidable. The principle used in correcting our description of the manifold correspondingly is that of the third law of duality: geometrical duality, which, to some extent, counteracts the necessary levelling effect of the other duality devices.

For the sustained description of a situation, or system of relations, we must choose a viewpoint. This introduces a certain arbitrariness in the description which, if carried to the bitter end, cannot but give a distorted picture of reality. The remedy is to change the viewpoint in the manner indicated by geometrical duality. In general, any punctual description requires correction by the corresponding primal description. Geometrical duality thus provides the limiting factor of the error element, and with it successive approximations to truth. This is how language takes care that trees do not grow into the sky.

Briefly, there are three limiting devices involved in language as an efficient code: for limiting

(a) the number of different code signs,
(b) the number of different code signs in use,
(c) the element of error,

each of which has the form of a law of duality.

In the same category of limiting devices in language belongs the use of a binary code and the inverse relation of symbolic length and probability according to Information-theory, which together limit the effort (costs) of transmitting information.

We may thus speak of a four-fold root of linguistic duality.

To summarise: According to whether we are concerned with the 'content' of linguistic forms, with these forms as 'expression' or with the relation between 'content' and 'expression', we distinguish three fields of the symbolic formulation of thought: symbolic logic, statistical linguistics, and linguistic duality in the geometrical sense of the term. The mathematics of each of these fields requires as a fundamental law a statement about absolute identity: of *meaning* (symbolic logic), of *expression* (statistical linguistics), of *relation* (linguistic duality as a case of geometrical duality). These statements are all "laws of duality", Boole's law and the statistical law being equations of the second degree, with the third law resembling the other two closely in its conceptual structure. We may say that the law of duality assumes a different shape according to whether it refers to single utterances or particular propositions (as in

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

symbolic logic), to language as a mass phenomenon (as in statistical linguistics) and to language as the linear expression of systematic and organised thought (as in works of literature).

This leads to the conclusion that the reason for the applicability of mathematics (algebra, statistics, geometry) to language, and through language to our description of the manifold, must be sought in the fact that the basic laws of thought and symbolic expression are intrinsically mathematical and dualistic. This view was put forward by the author in Vol. I, Sect. 15.1, 15.2, where the philosophical speculations which are known under the name of dualism are interpreted as manifestations of the geometrical variety of linguistic duality, with the consequence that such dual patterns as for instance the Hegelian-Marxian "dialectic movement of history" is recognised to be nothing but a hypostasised law of language. This view receives the strongest support through the fact that about 100 years ago the same idea occurred to G. Boole, but with regard to his algebraic law of duality. He says (p. 412 l.c.):

"In equally intimate alliance with that law of thought which is expressed by an equation of the second degree, and which has been termed in this treatise the law of duality, stands the tendency of ancient thought to those forms of philosophical speculation which are known under the name of dualism ... [3] The logical elements which underlie all these speculations, and from which they appear to borrow at least their form, it would be easy to trace in the outlines of more modern systems, more especially in that association of the doctrine of absolute unity with the distinction of the ego and non-ego as the type of Nature, which forms the basis of the philosophy of Hegel ... The wide prevalence of the particular theories which we have considered, together with their manifest analogy with the expressed laws of thought, may justly be conceived to indicate a connexion between the two systems ... The operation of that law of thought termed in this work the law of duality may have its own peculiar tendency to error, exalting mere want of agreement into contrariety, and thus form a world which we necessarily view as formed of parts supplemental to each other, framing the conception of a world fundamentally divided by opposing powers."

This recalls vividly the closing words of Eddington's *Space, Time and Gravitation* (1920):

"We have found that where science has progressed the farthest, the mind has but regained from nature that which the mind has put into nature ... We have found a strange foot-print on the shore of the unknown. We have devised profound theories, one after another, to account for its origin. At last, we have succeeded in reconstructing the creature that made the foot-print. And lo! it is our own."

[3] (Here follow illustrations from ancient philosophy.)

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

MATHEMATICAL APPENDICES

APPENDIX A

(1) If for the lognormal distribution with arithmetic mean μ and standard deviation σ we write $\Lambda(\mu, \sigma^2)$, the j'th moment distribution function of $\Lambda(\mu, \sigma^2)$ is defined by

$$\Lambda_j(x \mid \mu, \sigma^2) = \frac{\mathrm{I}}{\lambda'_j} \int_1^\infty u^j \, d\Lambda(u \mid \mu, \sigma^2), \tag{1}$$

and the fundamental theorem of the moment distribution is:

The j'th distribution of a lognormal variate with parameters μ and σ^2 is also a lognormal variate with parameters $\mu + j\sigma^2$ and σ^2 respectively.

Proof[1]:

$$\begin{aligned}
\Lambda_j(x \mid \mu, \sigma^2) &= \frac{\mathrm{I}}{\lambda'_j} \int_0^x u^j \, d\Lambda(u \mid \mu, \sigma^2) \\
&= e^{-j\mu - \frac{1}{2}j^2\sigma^2} \int_0^x e_j{}^{\log u} \frac{1}{u\sigma/(2\pi)} \exp\left\{-\frac{1}{2\sigma^2}(\log u - \mu)^2\right\} du \\
&= \int_0^x \frac{1}{u\sigma\sqrt{(2\pi)}} \exp\left\{-\frac{\mathrm{I}}{2\sigma^2}(\log u - \mu - j\sigma^2)^2\right\} du \\
&= \Lambda(x \mid \mu + j\sigma^2, \sigma^2)
\end{aligned} \tag{2}$$

(2) This simple result enables us to obtain an explicit expression for Gini's coefficient of mean difference:

The coefficient of mean difference of $\Lambda(\mu, \sigma^2)$ is

$$G = 2\left\{2N\left(\frac{\sigma}{2}\middle|0,1\right) - 1\right\}.$$

Proof[1]:

$$\begin{aligned}
G &= \int_0^\infty \int_0^\infty |u - v| \, d\Lambda(u) \, d\Lambda(v) \\
&= \int_0^\infty \int_0^\infty (u - v) \, d\Lambda(u) \, d\Lambda(v) + \int_0^\infty \int_u^\infty (v - u) \, d\Lambda(u) \, d\Lambda(v)
\end{aligned}$$

[1] Aitchison and Brown, 1957.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$= 2\int_0^\infty \int_0^u (u - v)\, d\Lambda(u)\, d\Lambda(v)$$

$$= 2\int_0^\infty u\Lambda(u)\, d\Lambda(u) - 2\int_0^\infty \int_0^u v\, d\Lambda(u)\, d\Lambda(v)$$

$$= 2\alpha\int_0^\infty \Lambda(u)\, d\Lambda_1(u) - 2\alpha \int_0^\infty \Lambda_1(u)\, d\Lambda(u)$$

where α is the arithmetic mean of the lognormal variate.

By the theorem about moment distributions of a lognormal variate we have

$$G = 2\alpha\int_0^\infty \Lambda(u \mid \mu, \sigma^2)\, d\Lambda(u \mid \mu + \sigma^2, \sigma^2) - 2\alpha \int_0^\infty \Lambda(u \mid \mu + \sigma^2, \sigma^2)\, d\Lambda(u \mid \mu, \sigma^2) \quad (5)$$

$$= 2\alpha\left\{\Lambda(1 \mid -\sigma^2, 2\sigma^2) - \Lambda(1 \mid \sigma^2, 2\sigma^2)\right\}. \quad (6)$$

using the reproductive property of lognormal variates.

Thus
$$G = 2\alpha\left\{2\Lambda(1 \mid -\sigma^2, 2\sigma^2) - 1\right\} \quad (7)$$

Considering now that if x is lognormally distributed, $y = \log x$ is normally distributed (for which we shall use the symbol N), we can express G in terms of N in respect of α, σ; if, moreover, we transform the resulting normal variate to one with zero mean and unit standard deviation and have

$$G = 2\alpha\left\{2N\left(\frac{\sigma}{\sqrt{2}} \,\middle|\, 0,1\right) - 1\right\}. \quad (8)$$

(3) The relation between G and the Lorenz factor L is as follows.

The formal definition of the Lorenz measure is

$$L = 1 - 2\int_0^\infty F_1(x)\, dF(x) \quad (9)$$

Using the theorem about moment distributions of lognormal variates, and substituting the explicit form for $F_1(x)$ and for $F(x)$ we have

$$\mathrm{L} = 1 - 2\int_0^\infty \Lambda(x \mid \mu + \sigma^2, \sigma^2)\, d\Lambda(x \mid \mu, \sigma^2) \quad (10)$$

$$= 1 - 2\Lambda(1 \mid \sigma^2, 2\sigma^2)$$

$$= 1 - 2N\left(\frac{-\sigma}{\sqrt{2}} \,\middle|\, 0,1\right)$$

$$= 2N\left(\frac{\sigma}{\sqrt{2}} \,\middle|\, 0,1\right) - 1. \quad (11)$$

This shows
$$G = 2\alpha L$$

or
$$L = G/2\alpha$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

In words: The Lorenz measure of concentration equals Gini's coefficient of mean difference divided by twice the arithmetic mean.

APPENDIX A1

LORENZ MEASURE AS A FUNCTION OF THE STANDARD DEVIATION[2]

σ	Lorenz measure $2N\left(\frac{\sigma}{\sqrt{2}}\middle\vert 0,1\right)-1$	σ	Lorenz measure $2N\left(\frac{\sigma}{\sqrt{2}}\middle\vert 0,1\right)-1$
0.0	0	1.6	0.7415
0.05	0.0279	1.7	0.7699
0.1	0.0566	1.8	0.7959
0.15	0.0844	1.9	0.8198
0.2	0.1121		
		2.0	0.8415
0.25	0.1405	2.1	0.8611
0.3	0.1679	2.2	0.8812
0.35	0.1951	2.3	0.8969
0.4	0.2228	2.4	0.9109
0.45	0.2495		
		2.5	0.9233
0.5	0.2767	2.6	0.9342
0.55	0.3027	2.7	0.9439
0.6	0.3284	2.8	0.9523
0.65	0.3545	2.9	0.9596
0.7	0.3794		
		3.0	0.9660
0.75	0.4039	3.1	0.9715
0.8	0.4286	3.2	0.9762
0.85	0.4522	3.3	0.9802
0.9	0.4752	3.4	0.9836
0.95	0.4984		
		3.5	0.9869
1.0	0.5204	3.6	0.9892
1.1	0.5646	3.7	0.9912
1.2	0.6047	3.8	0.9929
1.3	0.6424	3.9	0.9942
1.4	0.6778		
		4.0	0.9953
1.5	0.7108		

APPENDIX B

Definitions:

$$-\Sigma p_n \log_e p_n = H,\ \Sigma n p_n = M,\ \Sigma p_n = 1 \qquad (1)$$

If M is given, to maximise H.

[2] After Aitchison and Brown, 1957.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

We differentiate the three equations (1) partially with respect to p_n, and introduce Lagrange undetermined multipliers λ and μ. We get for all n,

$$1 + \log_e p_n + \lambda + n\mu = 0 \tag{2}$$

from which we deduce equ. 20 Chpt. XIII as follows:

$$p_n = e^{-N\mu - \lambda - 1} \tag{3}$$

Thus the p form a geometric series with $e^{-\mu}$ as common factor, and

$$p_1 = e^{-\mu - \lambda - 1}$$

Since for a geometric series $p_1 = 1/M$, and common factor $e^{-\mu} = 1 - p_1$

$$e^{-\mu - \lambda - 1} = 1 - e^{-\mu}, \tag{4}$$

furthermore, since

$$p_1 = e^{-\mu - \lambda - 1} = 1/M$$

and $$e^{-\mu} = 1 - 1/M,$$

we have $$e^{-\lambda - 1} = 1/M - 1$$

and $$p_n = 1/M\,(1 - 1/M)^{n-1}$$

$H_{\max}$ as the negative arithmetic mean of $\log_2 p_n$ is then calculated as

$$H_{\max} = -\Sigma\, p_n \log_2 p_n = \Sigma \frac{1}{M}\left(1 - \frac{1}{N}\right)^{n-1}\left\{-\log_2 M + (n-1)\log_2 \frac{M-1}{M}\right\}$$

which, after a little algebra, reduces to

$$M \log_2 M - (M-1)\log_2 (M-1) \tag{5}$$

APPENDIX C

$$K = \Sigma \frac{n_i\,(n_i - 1)}{N\,(N-1)} \tag{1}$$

where n_i is the number of occurrences of the vocabulary item i and N the total number of occurrences in the sample (text length).

We may write

$$\Sigma\,(n_i^2 - n_i) \approx \varkappa_p N^2 - \varkappa_p N; \tag{2}$$

$$\Sigma\, n_i = N\,; \tag{3}$$

$$\Sigma\, n_i^2 \approx \varkappa_p N^2 - \varkappa_p N + N \tag{4}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

If there is no significant difference between two text samples (having the same vocabulary) with regard to the frequency of use of words, the combined sample must again show the same frequency distribution of the vocabulary (the null-hypothesis).

Denoting the words in the first sample by a_{t_1} and their frequency by n_{t_1}, the words of the second sample by a_{t_2} and their frequency by n_{t_2}, and the total number of occurrences in these samples by N_1 and N_2, respectively, the combination gives

$$\Sigma({n_{t_1}}^2 + {n_{t_2}}^2) \approx \varkappa_p(N_1 + N_2)^2 - \varkappa_p(N_1 + N_2) + (N_1 + N_2) \tag{5}$$

$$\Sigma {n_{t_1}}^2 + \Sigma {n_{t_2}}^2 + 2\Sigma n_{t_1} n_{t_2} \approx \varkappa_p({N_1}^2 + {N_2}^2 + 2N_1 N_2) - \varkappa_p(N_1 + N_2) + (N_1 + N_2) \tag{6}$$

$$\Sigma {n_{t_1}}^2 \approx \varkappa_p {N_t}^2 - \varkappa_p N_1 + N_1 \tag{7}$$

$$\Sigma {n_{t_2}}^2 \approx \varkappa_p {N_2}^2 - \varkappa_p N_2 + N_2 \tag{8}$$

Substituting in (6), there remains

$$2\Sigma n_{t_1} n_{t_2} \approx \varkappa_p(2N_1 N_2) \tag{9}$$

$$\chi = \frac{\Sigma n_{t_1} n_{t_2}}{N_1 N_2} \approx \varkappa. \tag{10}$$

In words: if the two text samples are random samples from the same basic text, their combination according to formula (10) should equal the repeat rate, and for great samples K.[3]

APPENDIX D

We denote by c_m a measure of heterogeneity of the statistical population such that $c_m = {p_1}^m + {p_2}^m + \ldots {p_s}^m$, where $p\mu$ is the probability of occurrence of the μ-th word. Then the expectation of that parameter for which we will write c_m is given by:

$$c_m = \frac{1}{N^{[m]}} \sum_r r^{[m]} n_r = \frac{1}{N^{[m]}} \sum_r r^{[m]} \sum_{\mu=1}^{s} \binom{N}{r} {p_\mu}^r (r\, p_\mu)^{N-r} \tag{1}$$

$$= \sum_{\mu=1}^{s} \sum_r \frac{1}{N^{[m]}} r^{[m]} \binom{N}{r} {p_\mu}^r (1 - p_\mu)^{N-r}$$

$$= \sum_{\mu=1}^{s} \sum_r \frac{(N-m)!}{(r-m)!\,[N-m-(r-m)]!} {p_\mu}^{r-m} (1 - p_\mu)^{N-m-(r-m)} p_\mu^m$$

$$= \sum_{\mu=1}^{s} p_\mu^m [p_\mu + (1 - p_\mu)]^{N-m}$$

$$= \sum_{\mu=1}^{s} p_\mu^m = c_{m,0} \tag{2}$$

[3] M. de Vries, *Kwantifiicering in taal en cryptp-analyse*, Rapport SP 32, Statistische Afdeling, Mathem. Centrum (Amsterdam, 1954).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

In other words $\hat{c}_m$ is an unbiased estimate of c_m.

For our case $m = 2$, and we write

$$\hat{c}_2 = \frac{1}{N[^2]} \Sigma r\,(r-1)\,n_r = \frac{S_2 - N}{N\,(N-1)} \tag{3}$$

Yule's characteristic K is equal to $\hat{c}_2\,(1 - 1/N)$:

$$\hat{c}_2 \left(1 - \frac{1}{N}\right) = \frac{S_2 - N}{N\,(N-1)} \cdot \frac{N-1}{N} = \frac{S_2 - N}{N^2} = K \tag{4}$$

and thus tends to $p_1^2 + p_2^2 + \ldots$ as $N \to \infty$. It appears thus to be the repeat of the linguistic events.

APPENDIX E

THE VALUES OF $-3.322\ p_i \log p_i$ FOR VALUES OF p_i FROM .001 TO .999[1]

p_i	0	1	2	3	4	5	6	7	8	9
00	000	010	018	025	032	038	044	050	056	061
01	066	072	077	081	086	091	095	100	104	109
02	113	117	121	125	129	133	137	141	144	148
03	152	155	159	162	166	169	173	176	179	183
04	186	189	192	195	198	201	204	207	210	213
05	216	219	222	225	227	230	233	235	238	241
06	243	246	249	251	254	256	259	261	264	266
07	269	271	273	276	278	280	283	285	287	289
08	292	294	296	298	300	302	305	307	309	311
09	313	315	317	319	321	323	325	327	328	330
10	332	334	336	338	340	341	343	345	347	349
11	350	352	354	355	357	359	361	362	364	365
12	367	369	370	372	373	375	377	378	380	381
13	383	384	386	387	389	390	391	393	394	396
14	397	399	400	401	403	404	405	407	408	409
15	411	412	413	414	416	417	418	419	421	422
16	423	424	425	427	428	429	430	431	432	433
17	435	436	437	438	439	440	441	442	443	444
18	445	446	447	448	449	450	451	452	453	454
19	455	456	457	458	459	460	461	462	463	463
20	464	465	466	467	468	469	469	470	471	472
21	473	474	474	475	476	477	477	478	479	480
22	481	481	482	483	483	484	485	486	486	487
23	488	488	489	490	490	491	492	492	493	494
24	494	495	495	496	497	497	498	498	499	499
25	500	501	501	502	502	503	503	504	504	505
26	505	506	506	507	507	508	508	509	509	510
27	510	510	511	511	512	512	513	513	513	514
28	514	515	515	515	516	516	516	517	517	518
29	518	518	519	519	519	520	520	520	521	521
30	521	521	522	522	522	523	523	523	523	524
31	524	524	524	525	525	525	525	525	526	526

[1] Decimal point omitted before p_i and $-3.222\ p_i \log p_i$.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

p_t	0	1	2	3	4	5	6	7	8	9
32	526	526	526	527	527	527	527	527	528	528
33	528	528	528	528	528	529	529	529	529	529
34	529	529	529	530	530	530	530	530	530	530
35	530	530	530	530	530	530	530	530	531	531
36	531	531	531	531	531	531	531	531	531	531
37	531	531	531	531	531	531	531	531	531	531
38	530	530	530	530	530	530	530	530	530	530
39	530	530	530	530	529	529	529	529	529	529
40	529	529	529	528	528	528	528	528	528	528
41	527	527	527	527	527	527	527	526	526	526
42	526	525	525	525	525	525	525	524	524	524
43	524	523	523	523	523	522	522	522	522	521
44	521	521	521	520	520	520	520	519	519	519
45	518	518	518	517	517	517	517	516	516	516
46	515	515	515	514	514	514	513	513	513	512
47	512	512	511	511	511	510	510	509	509	509
48	508	508	508	507	507	506	506	506	505	505
49	504	504	503	503	503	502	502	501	505	500
50	500	500	499	499	498	498	497	497	496	496
51	495	495	495	494	494	493	493	492	492	491
52	491	490	490	489	489	488	488	487	486	486
53	485	485	484	484	483	483	482	482	481	481
54	480	479	479	478	478	477	477	476	475	475
55	474	474	473	473	472	471	471	470	470	469
56	468	468	467	467	466	465	465	464	463	463
57	462	462	461	460	460	459	458	458	457	456
58	456	455	455	454	453	453	452	451	450	450
59	449	448	448	447	446	446	445	444	444	443
60	442	442	441	440	439	439	438	437	436	436
61	435	434	434	433	432	431	431	430	429	428
62	428	427	426	425	425	424	423	422	421	421
63	420	419	418	418	417	416	415	414	414	413
64	412	411	410	410	409	408	407	406	406	405
65	404	403	402	401	401	400	399	398	397	397
66	396	395	394	393	392	391	391	390	389	389
67	387	386	385	384	384	383	382	381	380	379
68	378	377	377	376	375	374	373	372	371	370
69	369	368	368	367	366	365	364	363	362	361
70	360	359	358	357	356	356	355	354	353	352
71	351	350	349	348	347	346	345	344	343	342
72	341	340	339	338	337	336	335	334	333	332
73	331	330	329	328	327	326	325	324	323	322
74	321	320	319	318	317	316	315	314	313	312
75	311	310	309	308	307	306	305	304	303	302
76	301	300	299	298	297	296	295	294	292	291
77	290	289	288	287	286	285	284	283	282	281
78	280	279	277	276	275	274	273	272	271	270
79	269	268	266	265	264	263	262	261	260	259
80	258	256	255	254	253	252	251	250	249	247
81	246	245	244	243	242	241	239	238	237	236
82	235	234	232	231	230	229	228	227	225	224
83	223	222	221	219	218	217	216	215	214	212
84	211	210	209	208	206	205	204	203	202	200

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

p_t	0	1	2	3	4	5	6	7	8	9
85	199	198	197	196	194	193	192	191	190	188
86	187	186	185	183	182	181	180	179	177	176
87	175	174	172	171	170	169	167	166	165	164
88	162	161	160	159	157	156	155	153	152	151
89	150	148	147	146	145	143	142	141	139	138
90	137	136	134	133	132	130	129	128	126	125
91	124	123	121	120	119	117	116	115	113	112
92	111	109	108	107	105	104	103	101	100	099
93	097	096	095	093	092	091	089	088	087	085
94	084	083	081	080	079	077	076	074	073	072
95	070	069	068	066	065	063	062	061	059	058
96	057	054	053	051	051	050	048	047	045	044
97	043	041	040	038	037	036	034	033	031	030
98	029	027	026	024	023	021	020	019	017	016
99	014	013	011	010	009	007	006	004	003	001
100	000									

APPENDIX F

Values of the proportion of monotypes (f_1) and of the mean size of genus (M) for values of the time τ from 1 to 10 (the unit being the doubling period for species within the genus) and of the ratio ϱ (of the chance of a specific segmentation to the chance of a generic segmentation) from 1.0 to 3.0 (decimal point omitted before f_1)

Ratio of chance of specific segmentation to chance of generic segmentation.			Time									
			1	2	3	4	5	6	7	8	9	10
	1.0	f_1	62,500	53,125	50,781	50,195	50,049	50,012	50,003	50,001	50,000	50,000
		M	1.693	2.386	3.079	3.773	4.466	5.159	5.852	6.545	7.238	7.931
	1.1	f_1	61,566	51,333	48,608	47,882	47,689	47,638	47,624	47,620	47,619	47,619
		M	1.715	2.477	3.289	4.153	5.074	6.054	7.098	8.211	9.395	10.656
	1.2	f_1	60,761	49,750	46,660	45,793	45,549	45,481	45,462	45,457	45,455	45,455
		M	1.735	2.560	3.485	4.524	5.691	7.000	8.470	10.119	11.971	14.049
	1.3	f_1	60,060	48,343	44,905	43,897	43,601	43,514	43,489	43,481	43,479	43,479
		M	1.752	2.634	3.669	4.883	6.309	7.981	9.944	12.247	14.949	18.121
	1.4	f_1	59,444	47,084	43,318	42,170	41,820	41,713	41,681	41,671	41,668	41,667
		M	1.767	2.701	3.840	5.229	6.921	8.985	11.500	14.566	18.304	22.860
	1.5	f_1	58,899	45,953	41,875	40,591	40,186	40,059	40,018	40,006	40,002	40,001
		M	1.780	2.762	4.000	5.560	7.524	10.000	13.119	17.049	22.000	28.238
	1.6	f_1	58,413	44,930	40,559	39,141	38,682	38,533	38,485	38,469	38,464	38,462
		M	1.792	2.818	4.149	5.876	8.115	11.018	14.784	19.667	25.999	34.212
	1.7	f_1	57,977	44,001	39,353	37,807	37,293	37,122	37,065	37,046	37,040	37,038
		M	1.802	2.869	4.289	6.178	8.690	12.032	16.479	22.393	30.262	40.730

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Ratio of chance of specific segmentation to chance of generic segmentation.			Time									
			1	2	3	4	5	6	7	8	9	10
	1.8	f_1	57,584	43,154	38,245	36,575	36,007	35,814	35,748	35,726	35,718	35,716
		M	1.812	2.916	4.420	6.465	9.249	13.037	18.191	25.205	34.750	47.738
	1.9	f_1	57,228	42,379	37,224	35,434	34,813	34,597	34,523	34,497	34,488	34,484
		M	1.820	2.960	4.542	6.739	9.790	14.027	19.910	28.080	39.426	55.180
	2.0	f_1	56,904	41,667	36,280	34,375	33,702	33,464	33,379	33,350	33,339	33,335
		M	1.828	3.000	4.657	7.000	10.314	15.000	21.627	31.000	44.255	63.000
	2.1	f_1	56,607	41,010	35,404	33,389	32,664	32,404	32,311	32,277	32,265	32,261
		M	1.836	3.037	4.765	7.248	10.819	15.953	23.335	33.948	49.206	71.144
	2.2	f_1	56,335	40,403	34,590	32,468	31,695	31,412	31,309	31,272	31,258	31,253
		M	1.842	3.072	4.866	7.485	11.307	16.885	25.027	36.909	54.251	79.561
	2.3	f_1	56,084	39,840	33,831	31,608	30,786	30,482	30,369	30,327	30,312	30,306
		M	1.849	3.104	4.962	7.710	11.777	17.795	26.698	39.872	59.364	88.204
	2.4	f_1	55,852	39,316	33,122	30,801	29,932	29,607	29,485	29,439	29,422	29,416
		M	1.854	3.134	5.052	7.925	12.230	18.681	28.345	42.826	64.522	97.030
	2.5	f_1	55,638	38,828	32,458	30,044	29,129	28,783	28,652	28,602	28,583	28,576
		M	1.860	3.162	5.137	8.130	12.667	19.543	29.965	45.763	69.707	106.000
	2.6	f_1	55,438	38,372	31,835	29,332	28,373	28,006	27,865	27,811	27,791	27,783
		M	1.864	3.189	5.218	8.326	13.087	20.381	31.556	48.675	74.901	115.078
	2.7	f_1	55,252	37,944	31,250	28,660	27,659	27,271	27,122	27,064	27,041	27,032
		M	1.869	3.214	5.294	8.512	13.492	21.196	33.115	51.557	80.089	124.233
	2.8	f_1	55,079	37,544	30,699	28,027	26,984	26,576	26,418	26,356	26,331	26,322
		M	1.873	3.237	5.366	8.691	13.882	21.987	34.643	54.404	85.259	133.436
	2.9	f_1	54,916	37,167	30,179	27,428	26,344	25,918	25,750	25,684	25,658	25,648
		M	1.877	3.259	5.435	8.861	14.257	22.755	36.137	57.211	90.399	142.663
	3.0	f_1	54,764	36,812	29,687	26,860	25,738	25,293	25,116	25,046	25,018	25,007
		M	1.881	3.280	5.500	9.024	14.619	23.500	37.598	59.976	95.500	151.890

APPENDIX G

Table of the *Random Partitioning Function* giving the probability to six decimal places of a word with a particular frequency X occurring in K combinations of R text parts or samples.[2] Multiplication of the probability by the appropriate f_x gives the random partitioning numbers whose sum over all the values of is X the number of words common to K parts or samples.

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
$X = 1$						
1	1.000000	0.500000	0.333333	0.250000	0.200000	1
$X = 2$						
1	1.000000	0.250000	0.111111	0.062500	0.040000	1
2		0.500000	0.222222	0.125000	0.080000	2

[2] The significance level is 0.0001. Non-significant results are not printed, except if the number in the first column is significant, which is due to how the electronic computer was programmed.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
$X = 3$						
1	1.000000	0.125000	0.037037	0.015625	0.008000	1
2		0.750000	0.222222	0.093750	0.048000	2
3			0.222222	0.093750	0.048000	3
$X = 4$						
1	1.000000	0.062500	0.012346	0.003906	0.001600	1
2		0.875000	0.172840	0.054688	0.022400	2
3			0.444444	0.140625	0.057600	3
4				0.093750	0.038400	4
$X = 5$						
1	1.000000	0.031250	0.004115	0.000977	0.000320	1
2		0.937500	0.123457	0.029297	0.009600	2
3			0.617284	0.146484	0.048000	3
4				0.234375	0.076800	4
5					0.038400	5
$X = 6$						
1	1.000000	0.015625	0.001372	0.000244	0.000064	1
2		0.968750	0.085048	0.015137	0.003968	2
3			0.740741	0.131836	0.034560	3
4				0.380859	0.099840	4
5					0.115200	5
$X = 7$						
1	1.000000	0.007813	0.000457	0.000061	0.000013	1
2		0.984375	0.057613	0.007690	0.001613	2
3			0.825789	0.110229	0.023117	3
4				0.512695	0.107520	4
5					0.215040	5
$X = 8$						
1	1.000000	0.003906	0.000152	0.000015	0.000003	1
2		0.992188	0.038714	0.003876	0.000650	2
3			0.883402	0.088440	0.014838	3
4				0.622925	0.104509	4
5					0.322560	5
$X = 9$						
1	1.000000	0.001953	0.000051	0.000004	0.000001	1
2		0.996094	0.025911	0.001945	0.000261	2
3			0.922116	0.069237	0.009293	3
4				0.711365	0.095478	4
5					0.427069	5
$X = 10$						
1	1.000000	0.000977	0.000017	0.000001	0.000000	1
2		0.998047	0.017308	0.000975	0.000105	2
3			0.948026	0.053387	0.005732	3
4				0.780602	0.083816	4
5					0.522547	5

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	K
$X=11$						
1	1.000000	0.000488	0.000006	0.000000	0.000000	1
2		0.999023	0.011550	0.000488	0.000042	2
3			0.965334	0.040771	0.003502	3
4				0.833988	0.071639	4
5					0.606364	5
$X=12$						
1	1.000000	0.000244	0.000002	0.000000	0.000000	1
2		0.999512	0.007704	0.000244	0.000017	2
3			0.976884	0.030944	0.002126	3
4				0.874759	0.060113	4
5					0.678003	5
$X=13$						
1	1.000000	0.000122	0.000001	0.000000	0.000000	1
2		0.999756	0.005137	0.000122	0.000007	2
3			0.984587	0.023391	0.001286	3
4				0.905703	0.049792	4
5					0.738116	5
$X=14$						
1	1.000000	0.000061	0.000000	0.000000	0.000000	1
2		0.999878	0.003425	0.000061	0.000003	2
3			0.989724	0.017635	0.000776	3
4				0.929094	0.040862	4
5					0.787907	5
$X=15$						
1	1.000000	0.000031	0.000000	0.000000	0.000000	1
2		0.999939	0.002284	0.000031	0.000001	2
3			0.993149	0.013272	0.000467	3
4				0.946729	0.033310	4
5					0.828769	5
$X=16$						
1	1.000000	0.000015	0.000000	0.000000	0.000000	1
2		0.999969	0.001522	0.000015	0.000000	2
3			0.995433	0.009977	0.000281	3
4				0.960001	0.027022	4
5					0.862079	5
$X=17$						
1	1.000000	0.000008	0.000000	0.000000	0.000000	1
2		0.999985	0.001015	0.000008	0.000000	2
3			0.996955	0.007494	0.000169	3
4				0.969978	0.021842	4
5					0.889101	5
$X=18$						
1	1.000000	0.000004	0.000000	0.000000	0.000000	1
2		0.999992	0.000677	0.000004	0.000000	2
3			0.997970	0.005626	0.000101	3

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
4				0.977472	0.017609	4
5					0.910943	5
$X = 19$						
1	1.000000	0.000002	0.000000	0.000000	0.000000	1
2		0.999996	0.000451	0.000002	0.000000	2
3			0.998647	0.004223	0.000061	3
4				0.983098	0.014168	4
5					0.928551	5
$X = 20$						
1	1.000000	0.000001	0.000000	0.000000	0.000000	1
2		0.999998	0.000301	0.000001	0.000000	2
3			0.999098	0.003168	0.000037	3
4				0.987321	0.011383	4
5					0.942719	5
$X = 21$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		0.999999	0.000200	0.000000	0.000000	2
3			0.999399	0.002377	0.000022	3
4				0.990489	0.009136	4
5					0.954102	5
$X = 22$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000134	0.000000	0.000000	2
3			0.999599	0.001783	0.000013	3
4				0.992866	0.007326	4
5					0.963238	5
$X = 23$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000089	0.000000	0.000000	2
3			0.999733	0.001337	0.000008	3
4				0.994649	0.005871	4
5					0.970564	5
$X = 24$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000059	0.000000	0.000000	2
3			0.999822	0.001003	0.000005	3
4				0.995987	0.004703	4
5					0.976436	5
$X = 25$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000040	0.000000	0.000000	2
3			0.999881	0.000752	0.000003	3
4				0.996990	0.003767	4
5					0.981139	5
$X = 26$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000026	0.000000	0.000000	2
3			0.999921	0.000564	0.000002	3
4				0.997742	0.003015	4
5					0.984905	5

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
$X = 27$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000018	0.000000	0.000000	2
3			0.999947	0.000423	0.000001	3
4				0.998307	0.002414	4
5					0.987921	5
$X = 28$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000012	0.000000	0.000000	2
3			0.999965	0.000317	0.000001	3
4				0.998730	0.001932	4
5					0.990335	5
$X = 29$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000008	0.000000	0.000000	2
3			0.999977	0.000238	0.000000	3
4				0.999048	0.001546	4
5					0.992267	5
$X = 30$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000005	0.000000	0.000000	2
3			0.999984	0.000179	0.000000	3
4				0.999286	0.001237	4
5					0.993813	5
$X = 31$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000003	0.000000	0.000000	2
3			0.999990	0.000134	0.000000	3
4				0.999464	0.000990	4
5					0.995050	5
$X = 32$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000002	0.000000	0.000000	2
3			0.999993	0.000100	0.000000	3
4				0.999598	0.000792	4
5					0.996039	5
$X = 33$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000002	0.000000	0.000000	2
3			0.999995	0.000075	0.000000	3
4				0.999699	0.000634	4
5					0.996831	5
$X = 34$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000001	0.000000	0.000000	2
3			0.999997	0.000057	0.000000	3
4				0.999774	0.000507	4
5					0.997465	5

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	K
$X=35$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000001	0.000000	0.000000	2
3			0.999998	0.000042	0.000000	3
4				0.999830	0.000406	4
5					0.997972	5
$X=36$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			0.999999	0.000032	0.000000	3
4				0.998773	0.000324	4
5					0.998378	5
$X=37$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			0.999999	0.000024	0.000000	3
4				0.999905	0.000260	4
5					0.998702	5
$X=38$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			0.999999	0.000018	0.000000	3
4				0.999928	0.000208	4
5					0.998962	5
$X=39$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000013	0.000000	3
4				0.999946	0.000166	4
5					0.999169	5
$X=40$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000010	0.000000	3
4				0.999960	0.000133	4
5					0.999335	5
$X=41$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000008	0.000000	3
4				0.999970	0.000106	4
5					0.999468	5
$X=42$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000006	0.040000	3
4				0.999977	0.000085	4
5					0.999575	5
$X=43$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R=1$	$R=2$	$R=3$	$R=4$	$R=5$	K
3			1.000000	0.000004	0.000000	3
4				0.999983	0.000068	4
5					0.999660	5
$X=44$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000003	0.000000	3
4				0.999987	0.000054	4
5					0.999728	5
$X=45$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000002	0.000000	3
4				0.999990	0.000044	4
5					0.999782	5
$X=46$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000002	0.000000	3
4				0.999993	0.000035	4
5					0.999826	5
$X=47$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000001	0.000000	3
4				0.999995	0.000028	4
5					0.999861	5
$X=48$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000001	0.000000	3
4				0.999996	0.000022	4
5					0.999888	5
$X=49$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000001	0.000000	3
4				0.999997	0.000018	4
5					0.999911	5
$X=50$						
1	1,000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000001	0.000000	3
4				0.999998	0.000014	4
5					0.999929	5
$X=51$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
4				0.999998	0.000011	4
5					0.999943	5
$X = 52$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				0.999999	0.000009	4
5					0.999954	5
$X = 53$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				0.999999	0.000007	4
5					0.999963	5
$X = 54$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				0.999999	0.000006	4
5					0.999971	5
$X = 55$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				0.999999	0.000005	4
5					0.999977	5
$X = 56$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.0000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000004	4
5					0.999981	5
$X = 57$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000003	4
5					0.999985	5
$X = 58$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000002	4
5					0.999988	5
$X = 59$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000002	4
5					0.999990	5

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	K
$X = 60$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000002	4
5					0.999992	5
$X = 61$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000001	4
5					0.999994	5
$X = 62$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000001	4
5					0.999995	5
$X = 63$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000001	4
5					0.999996	5
$X = 64$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000001	4
5					0.999997	5
$X = 65$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000001	4
5					0.999997	5
$X = 66$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000000	4
5					0.999998	5
$X = 67$						
1	1.000000	0.000000	0.000000	0.000000	0.000000	1
2		1.000000	0.000000	0.000000	0.000000	2
3			1.000000	0.000000	0.000000	3
4				1.000000	0.000000	4
5					0.999998	5

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 1$						
1	0.166667	0.142857	0.125000	0.111111	0.100000	1
$X = 2$						
1	0.027778	0.020408	0.015625	0.012346	0.010000	1
2	0.055556	0.040816	0.031250	0.024691	0.020000	2
$X = 3$						
1	0.004630	0.002915	0.001953	0.001372	0.001000	1
2	0.027778	0.017493	0.011719	0.008230	0.006000	2
3	0.027778	0.017493	0.011719	0.008230	0.006000	3
$X = 4$						
1	0.000772	0.000416	0.000244	0.000152	0.000100	1
2	0.010802	0.005831	0.003418	0.002134	0.001400	2
3	0.027778	0.014994	0.008789	0.005487	0.003600	3
4	0.018519	0.009996	0.005859	0.003658	0.002400	4
$X = 5$						
1	0.000129	0.000059	0.000031	0.000017	0.000010	1
2	0.003858	0.001785	0.000916	0.000508	0.000300	2
3	0.019290	0.008925	0.004578	0.002540	0.001500	3
4	0.030864	0.014280	0.007324	0.004064	0.002400	4
5	0.015432	0.007140	0.003662	0.002032	0.001200	5
$X = 6$						
2	0.001329	0.000527	0.000237	0.000117	0.000062	2
3	0.011574	0.004590	0.002060	0.001016	0.000540	3
4	0.033436	0.013260	0.005951	0.002935	0.001560	4
5	0.038580	0.015300	0.006866	0.003387	0.001800	5
6	0.015432	0.006120	0.002747	0.001355	0.000720	6
$X = 7$						
2	0.000450	0.000153	0.000060	0.000026	0.000013	2
3	0.006451	0.002193	0.000861	0.000378	0.000181	3
4	0.030007	0.010200	0.004005	0.001756	0.000840	4
5	0.060014	0.020400	0.008011	0.003512	0.001680	5
6	0.054012	0.018360	0.007210	0.003161	0.001512	6
7		0.006120	0.002403	0.001054	0.000504	7
$X = 8$						
3	0.003451	0.001005	0.000345	0.000135	0.000058	3
4	0.024306	0.007082	0.002433	0.000948	0.000408	4
5	0.075017	0.021857	0.007510	0.002927	0.001260	5
6	0.114026	0.033222	0.011415	0.004449	0.001915	6
7		0.024480	0.008411	0.003278	0.001411	7
8			0.002403	0.000937	0.000403	8
$X = 9$						
3	0.001801	0.000450	0.000135	0.000047	0.000018	3
4	0.018504	0.004621	0.001389	0.000481	0.000186	4
5	0.082769	0.020670	0.006215	0.002153	0.000834	5
6	0.189043	0.047211	0.014194	0.004917	0.001905	6
7		0.057702	0.017349	0.006010	0.002328	7
8			0.010815	0.003747	0.001452	8
9				0.000937	0.000363	9

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0. Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 10$						
3	0.000926	0.000198	0.000052	0.000016	0.000006	3
4	0.013537	0.002898	0.000762	0.000235	0.000082	4
5	0.084394	0.018065	0.004753	0.001464	0.000510	5
6	0.271812	0.058184	0.015307	0.004714	0.001644	6
7		0.104913	0.027600	0.008499	0.002964	7
8			0.028163	0.008673	0.003024	8
9				0.004683	0.001633	9
10					0.000363	10
$X = 11$						
3	0.000471	0.000086	0.000020	0.000005	0.000002	3
4	0.009642	0.001769	0.000407	0.000111	0.000035	4
5	0.081609	0.014974	0.003447	0.000943	0.000296	5
6	0.356206	0.065356	0.015044	0.004118	0.001292	6
7		0.163096	0.037543	0.010277	0.003225	7
8			0.055763	0.015264	0.004790	8
9				0.013356	0.004191	9
10					0.001996	10
$X = 12$						
4	0.006742	0.001060	0.000214	0.000052	0.000015	4
5	0.076043	0.011959	0.002409	0.000586	0.000166	5
6	0.437816	0.068854	0.013868	0.003374	0.000953	6
7		0.228452	0.046014	0.011196	0.003162	7
8			0.093306	0.022703	0.006412	8
9				0.028620	0.008083	9
10					0.006187	10
$X = 13$						
4	0.004654	0.000627	0.000111	0.000024	0.000006	4
5	0.068987	0.009300	0.001639	0.000354	0.000090	5
6	0.513858	0.069268	0.012208	0.002640	0.000671	6
7		0.297307	0.052397	0.011333	0.002881	7
8			0.139321	0.030132	0.007659	8
9				0.051323	0.013046	9
10					0.014270	10
$X = 14$						
4	0.003183	0.000368	0.000057	0.000011	0.000002	4
5	0.061367	0.007091	0.001093	0.000210	0.000048	5
6	0.582845	0.067344	0.010385	0.001997	0.000457	6
7		0.366575	0.056530	0.010868	0.002486	7
8			0.191718	0.036858	0.008432	8
9				0.081455	0.018634	9
10					0.027316	10
$X = 15$						
4	0.002162	0.000214	0.000029	0.000005	0.000001	4
5	0.053792	0.005327	0.000719	0.000123	0.000025	5
6	0.644213	0.063801	0.008609	0.001471	0.000303	6
7		0.433919	0.058550	0.010006	0.002060	7
8			0.248248	0.042423	0.008734	8
9				0.118313	0.024360	9
10					0.045950	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 16$						
4	0.001462	0.000124	0.000015	0.000002	0.000000	4
5	0.046628	0.003958	0.000467	0.000071	0.000013	5
6	0.698004	0.059253	0.006996	0.001063	0.000197	6
7		0.497720	0.058764	0.008926	0.001654	7
8			0.306798	0.046603	0.008636	8
9				0.160736	0.029785	9
10					0.070310	10
$X = 17$						
4	0.000984	0.000072	0.000007	0.000001	0.000000	4
5	0.040075	0.002916	0.000301	0.000041	0.000007	5
6	0.744632	0.054181	0.005597	0.000756	0.000126	6
7		0.556973	0.057540	0.007769	0.001296	7
8			0.365562	0.049359	0.008232	8
9				0.207338	0.034578	9
10					0.100094	10
$X = 18$						
4	0.000661	0.000041	0.000004	0.000000	0.000000	4
5	0.034216	0.002134	0.000193	0.000023	0.000003	5
6	0.784707	0.048940	0.004424	0.000531	0.000080	6
7		0.611154	0.055245	0.006631	0.000995	7
8			0.423102	0.050781	0.007622	8
9				0.256698	0.038529	9
10					0.134673	10
$X = 19$						
4	0.000443	0.000024	0.000002	0.000000	0.000000	4
5	0.029064	0.001554	0.000123	0.000013	0.000002	5
6	0.818923	0.043778	0.003463	0.000369	0.000050	6
7		0.660094	0.052211	0.005570	0.000752	7
8			0.478348	0.051032	0.006894	8
9				0.307478	0.041536	9
10					0.173202	10
$X = 20$						
5	0.024590	0.001127	0.000078	0.000007	0.000001	5
6	0.847988	0.038856	0.002689	0.000255	0.000031	6
7		0.703872	0.048714	0.004620	0.000562	7
8			0.530558	0.050313	0.006117	8
9				0.358511	0.043587	9
10					0.214737	10
$X = 21$						
5	0.020739	0.000815	0.000049	0.000004	0.000000	5
6	0.872577	0.034271	0.002075	0.000175	0.000019	6
7		0.742727	0.044978	0.003791	0.000415	7
8			0.579722	0.048829	0.005343	8
9				0.408824	0.044733	9
10					0.258324	10
$X = 22$						
5	0.017448	0.000587	0.000031	0.000002	0.000000	5
6	0.893317	0.030073	0.001594	0.000119	0.000012	6
7		0.776998	0.041172	0.003085	0.000304	7

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
8			0.624250	0.046774	0.004606	8
9				0.457653	0.045068	9
10					0.303057	10
$X = 23$						
5	0.014651	0.000423	0.000020	0.000001	0.000000	5
6	0.910765	0.026280	0.001218	0.000081	0.000007	6
7		0.807071	0.037419	0.002492	0.000221	7
8			0.665422	0.044319	0.003928	8
9				0.504427	0.044707	9
10					0.348125	10
$X = 24$						
5	0.012283	0.000304	0.000012	0.000001	0.000000	5
6	0.925415	0.022888	0.000929	0.000055	0.000004	6
7		0.833351	0.033808	0.002002	0.000160	7
8			0.702841	0.041610	0.003319	8
9				0.548746	0.043772	9
10					0.392832	10
$X = 25$						
5	0.010285	0.000218	0.000008	0.000000	0.000000	5
6	0.937698	0.019879	0.000706	0.000037	0.000003	6
7		0.856239	0.030395	0.001599	0.000115	7
8			0.736649	0.038766	0.002783	8
9				0.590356	0.042382	9
10					0.436604	10
$X = 26$						
5	0.008604	0.000156	0.000005	0.000000	0.000000	5
6	0.947983	0.017226	0.000535	0.000025	0.000002	6
7		0.876118	0.027213	0.001273	0.000082	7
8			0.767044	0.035880	0.002318	8
9				0.629122	0.040648	9
10					0.478985	10
$X = 27$						
5	0.007192	0.000112	0.000003	0.000000	0.000000	5
6	0.956586	0.014899	0.000405	0.000017	0.000001	6
7		0.893344	0.024279	0.001010	0.000059	7
8			0.794257	0.033025	0.001920	8
9				0.665002	0.038670	9
10					0.519634	10
$X = 28$						
5	0.006008	0.000080	0.000002	0.000000	0.000000	5
6	0.963778	0.012867	0.000306	0.000011	0.000001	6
7		0.908243	0.021599	0.000798	0.000042	7
8			0.818536	0.030253	0.001583	8
9				0.698027	0.036531	9
10					0.558303	10
$X = 29$						
5	0.005016	0.000057	0.000001	0.000000	0.000000	5
6	0.969786	0.011097	0.000231	0.000008	0.000000	6
7		0.921110	0.019167	0.000630	0.000030	7
8			0.840135	0.027601	0.001300	8

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
9				0.728280	0.034303	9
10					0.594834	10
$X = 30$						
5	0.004187	0.000041	0.000001	0.000000	0.000000	5
6	0.974802	0.009561	0.000174	0.000005	0.000000	6
7		0.932207	0.016973	0.000496	0.000021	7
8			0.859301	0.025094	0.001064	8
9				0.755881	0.032043	9
10					0.629137	10
$X = 31$						
5	0.003493	0.000029	0.000000	0.000000	0.000000	5
6	0.978989	0.008231	0.000131	0.000003	0.000000	6
7		0.941769	0.015004	0.000389	0.000015	7
8			0.876274	0.022746	0.000868	8
9				0.780975	0.029796	9
10					0.661180	10
$X = 32$						
5	0.002914	0.000021	0.000000	0.000000	0.000000	5
6	0.982482	0.007080	0.000099	0.000002	0.000000	5
7		0.949999	0.013243	0.000306	0.000010	7
8			0.891278	0.020565	0.000706	8
9				0.803721	0.027597	9
10					0.690976	10
$X = 33$						
5	0.002430	0.000015	0.000000	0.000000	0.000000	5
6	0.985396	0.006086	0.000074	0.000002	0.000000	6
7		0.957079	0.011674	0.000239	0.000007	7
8			0.904520	0.018552	0.000573	8
9				0.824286	0.025473	9
10					0.718573	10
$X = 34$						
5	0.002026	0.000011	0.000000	0.000000	0.000000	5
6	0.987826	0.005230	0.000056	0.000001	0.000000	6
7		0.963166	0.010280	0.000187	0.000005	7
8			0.916194	0.016703	0.000465	8
9				0.842838	0.023442	9
10					0.744046	10
$X = 35$						
5	0.001690	0.000008	0.000000	0.000000	0.000000	5
6	0.989852	0.004492	0.000042	0.000001	0.000000	6
7		0.968395	0.009043	0.000147	0.000004	7
8			0.926474	0.015014	0.000376	8
9				0.859541	0.021516	9
10					0.767488	10
$X = 36$						
5	0.001409	0.000005	0.000000	0.000000	0.000000	5
6	0.991542	0.003857	0.000032	0.000000	0.000000	6
7		0.972887	0.007950	0.000115	0.000003	7
8			0.935517	0.013476	0.000304	8

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
9				0.874555	0.019702	9
10					0.789003	10
$X = 37$						
5	0.001174	0.000004	0.000000	0.000000	0.000000	5
6	0.992950	0.003311	0.000024	0.000000	0.000000	6
7		0.976744	0.006984	0.000089	0.000002	7
8			0.943467	0.012080	0.000245	8
9				0.888031	0.018005	9
10					0.808706	10
$X = 38$						
5	0.000979	0.000003	0.000000	0.000000	0.000000	5
6	0.994125	0.002841	0.000018	0.000000	0.000000	6
7		0.980055	0.006131	0.000070	0.000001	7
8			0.950451	0.010818	0.000197	8
9				0.900112	0.016425	9
10					0.826711	10
$X = 39$						
5	0.000816	0.000002	0.000000	0.000000	0.000000	5
6	0.995103	0.002437	0.000013	0.000000	0.000000	6
7		0.982896	0.005380	0.000054	0.000001	7
8			0.956582	0.009678	0.000159	8
9				0.910929	0.014960	9
10					0.843136	10
$X = 40$						
5	0.000680	0.000001	0.000000	0.000000	0.000000	5
6	0.995919	0.002091	0.000010	0.000000	0.000000	6
7		0.985333	0.004720	0.000042	0.000001	7
8			0.961962	0.008651	0.000128	8
9				0.920607	0.013607	9
10					0.858097	10
$X = 41$						
5	0.000567	0.000001	0.000000	0.000000	0.000000	5
6	0.996599	0.001794	0.000008	0.000000	0.000000	6
7		0.987424	0.004138	0.000033	0.000000	7
8			0.966682	0.007727	0.000103	8
9				0.929258	0.012362	9
10					0.871704	10
$X = 42$						
5	0.000472	0.000001	0.000000	0.000000	0.000000	5
6	0.997166	0.001538	0.000006	0.000000	0.000000	6
7		0.989218	0.003628	0.000026	0.000000	7
8			0.970820	0.006898	0.000083	8
9				0.936985	0.011218	9
10					0.884066	10
$X = 43$						
6	0.997638	0.001319	0.000004	0.000000	0.000000	6
7		0.990756	0.003179	0.000020	0.000000	7
8			0.974448	0.006155	0.000066	8
9				0.943883	0.010171	9
10					0.895284	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 44$						
6	0.998032	0.001131	0.000003	0.000000	0.000000	6
7		0.992075	0.002785	0.000016	0.000000	7
8			0.977627	0.005489	0.000053	8
9				0.950038	0.009213	9
10					0.905454	10
$X = 45$						
6	0.998360	0.000970	0.000002	0.000000	0.000000	6
7		0.993206	0.002440	0.000012	0.000000	7
8			0.980413	0.004893	0.000043	8
9				0.955527	0.008340	9
10					0.914668	10
$X = 46$						
6	0.998633	0.000831	0.000002	0.000000	0.000000	6
7		0.994176	0.002137	0.000009	0.000000	7
8			0.982853	0.004360	0.000034	8
9				0.960419	0.007544	9
10					0.923007	10
$X = 47$						
6	0.998861	0.000713	0.000001	0.000000	0.000000	6
7		0.995007	0.001872	0.000007	0.000000	7
8			0.984990	0.003884	0.000027	8
9				0.964779	0.006821	9
10					0.930552	10
$X = 48$						
6	0.999051	0.000611	0.000001	0.000000	0.000000	6
7		0.995720	0.001639	0.000006	0.000000	7
8			0.986861	0.003459	0.000022	8
9				0.968663	0.006163	9
10					0.937372	10
$X = 49$						
6	0.999209	0.000524	0.000001	0.000000	0.000000	6
7		0.996331	0.001435	0.000004	0.000000	7
8			0.988500	0.003080	0.000018	8
9				0.972122	0.005567	9
10					0.943536	10
$X = 50$						
6	0.999341	0.000449	0.000001	0.000000	0.000000	6
7		0.996855	0.001256	0.000003	0.000000	7
8			0.989935	0.002741	0.000014	8
9				0.975202	0.005026	9
10					0.949102	10
$X = 51$						
6	0.999451	0.000385	0.000000	0.000000	0.000000	6
7		0.997304	0.001100	0.000003	0.000000	7
8			0.991191	0.002440	0.000011	8
9				0.977943	0.004536	9
10					0.954128	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0. Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 52$						
6	0.999542	0.000330	0.000000	0.000000	0.000000	6
7		0.997689	0.000963	0.000002	0.000000	7
8			0.992291	0.002171	0.000009	8
9				0.980383	0.004093	9
10					0.958664	10
$X = 53$						
6	0.999618	0.000283	0.000000	0.000000	0.000000	6
7		0.998019	0.000842	0.000002	0.000000	7
8			0.993253	0.001932	0.000007	8
9				0.982554	0.003692	9
10					0.962757	10
$X = 54$						
6	0.999682	0.000243	0.000000	0.000000	0.000000	6
7		0.998302	0.000737	0.000001	0.000000	7
8			0.994096	0.001719	0.000006	8
9				0.984486	0.003329	9
10					0.966449	10
$X = 55$						
6	0.999735	0.000208	0.000000	0.000000	0.000000	6
7		0.998545	0.000645	0.000001	0.000000	7
8			0.994833	0.001529	0.000005	8
9				0.986205	0.003001	9
10					0.969778	10
$X = 56$						
6	0.999779	0.000178	0.000000	0.000000	0.000000	6
7		0.998753	0.000565	0.000001	0.000000	7
8			0.995479	0.001360	0.000004	8
9				0.987734	0.002705	9
10					0.972779	10
$X = 57$						
6	0.999816	0.000153	0.000000	0.000000	0.000000	6
7		0.998931	0.000494	0.000001	0.000000	7
8			0.996043	0.001209	0.000003	8
9				0.989093	0.002438	9
10					0.975484	10
$X = 58$						
6	0.999847	0.000131	0.000000	0.000000	0.000000	6
7		0.999083	0.000433	0.000000	0.000000	7
8			0.996538	0.001076	0.000002	8
9				0.990303	0.002197	9
10					0.977922	10
$X = 59$						
6	0.999872	0.000112	0.000000	0.000000	0.000000	6
7		0.999214	0.000379	0.000000	0.000000	7
8			0.996970	0.000956	0.000002	8
9				0.991378	0.001979	9
10					0.980119	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 60$						
6	0.999894	0.000096	0.000000	0.000000	0.000000	6
7		0.999327	0.000331	0.000000	0.000000	7
8			0.997349	0.000851	0.000002	8
9				0.992335	0.001783	9
10					0.982099	10
$X = 61$						
6	0.999911	0.000082	0.000000	0.000000	0.000000	6
7		0.999423	0.000290	0.000000	0.000000	7
8			0.997680	0.000756	0.000001	8
9				0.993185	0.001606	9
10					0.983882	10
$X = 62$						
6	0.999926	0.000071	0.000000	0.000000	0.000000	6
7		0.999505	0.000254	0.000000	0.000000	7
8			0.997970	0.000672	0.000001	8
9				0.993942	0.001447	9
10					0.985488	10
$X = 63$						
6	0.999938	0.000061	0.000000	0.000000	0.000000	6
7		0.999576	0.000222	0.000000	0.000000	7
8			0.998224	0.000598	0.000001	8
9				0.994614	0.001303	9
10					0.986935	10
$X = 64$						
6	0.999949	0.000052	0.000000	0.000000	0.000000	6
7		0.999637	0.000194	0.000000	0.000000	7
8			0.998446	0.000532	0.000001	8
9				0.995212	0.001173	9
10					0.988238	10
$X = 65$						
6	0.999957	0.000045	0.000000	0.000000	0.000000	6
7		0.999688	0.000170	0.000000	0.000000	7
8			0.998640	0.000473	0.000001	8
9				0.995744	0.001057	9
10					0.989411	10
$X = 66$						
6	0.999964	0 000038	0.000000	0.000000	0.000000	6
7		0.999733	0.000149	0.000000	0.000000	7
8			0.998810	0.000420	0.000000	8
9				0.996216	0.000951	9
10					0.990468	10
$X = 67$						
6	0.999970	0.000033	0.000000	0.000000	0.000000	6
7		0.999771	0.000130	0.000000	0.000000	7
8			0.998959	0.000374	0.000000	8
9				0.996636	0.000857	9
10					0.991419	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 68$						
6	0.999975	0.000028	0.000000	0.000000	0.000000	6
7		0.999804	0.000114	0.000000	0.000000	7
8			0.999089	0.000332	0.000000	8
9				0.997010	0.000771	9
10					0.992276	10
$X = 69$						
6	0.999979	0.000024	0.000000	0.000000	0.000000	6
7		0.999832	0.000100	0.000000	0.000000	7
8			0.999203	0.000295	0.000000	8
9				0.997342	0.000694	9
10					0.993047	10
$X = 70$						
6	0.999983	0.000021	0.000000	0.000000	0.000000	6
7		0.999856	0.000087	0.000000	0.000000	7
8			0.999302	0.000262	0.000000	8
9				0.997637	0.000625	9
10					0.993742	10
$X = 71$						
6	0.999986	0.000018	0.000000	0.000000	0.000000	6
7		0.999876	0.000076	0.000000	0.000000	7
8			0.999390	0.000233	0.000000	8
9				0.997900	0.000563	9
10					0.994367	10
$X = 72$						
6	0.999988	0.000015	0.000000	0.000000	0.000000	6
7		0.999894	0.000067	0.000000	0.000000	7
8			0.999466	0.000207	0.000000	8
9				0.998133	0.000507	9
10					0.994929	10
$X = 73$						
6	0.999990	0.000013	0.000000	0.000000	0.000000	6
7		0.999909	0.000058	0.000000	0.000000	7
8			0.999533	0.000184	0.000000	8
9				0.998340	0.000456	9
10					0.995436	10
$X = 74$						
6	0.999992	0.000011	0.000000	0.000000	0.000000	6
7		0.999922	0.000051	0.000000	0.000000	7
8			0.999591	0.000164	0.000000	8
9				0.998525	0.000410	9
10					0.995892	10
$X = 75$						
6	0.999993	0.000010	0.000000	0.000000	0.000000	6
7		0.999933	0.000045	0.000000	0.000000	7
8			0.999642	0.000146	0.000000	8
9				0.998689	0.000370	9
10					0.996303	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 76$						
6	0.999994	0.000008	0.000000	0.000000	0.000000	6
7		0.999943	0.000039	0.000000	0.000000	7
8			0.999687	0.000130	0.000000	8
9				0.998834	0.000333	9
10					0.996672	10
$X = 77$						
6	0.999995	0.000007	0.000000	0.000000	0.000000	6
7		0.999951	0.000034	0.000000	0.000000	7
8			0.999726	0.000115	0.000000	8
9				0.998964	0.000299	9
10					0.997005	10
$X = 78$						
6	0.999996	0.000006	0.000000	0.000000	0.000000	6
7		0 999958	0 000030	0 000000	0.000000	7
8			0.999760	0.000102	0.000000	8
9				0.999079	0.000269	9
10					0.997304	10
$X = 79$						
6	0.999997	0.000005	0.000000	0.000000	0.000000	6
7		0.999964	0.000026	0.000000	0.000000	7
8			0.999790	0.000091	0.000000	8
9				0.999181	0.000243	9
10					0.997573	10
$X = 80$						
6	0.999997	0.000004	0.000000	0.000000	0.000000	6
7		0.999969	0.000023	0.000000	0.000000	7
8			0.999816	0.000081	0.000000	8
9				0.999272	0.000218	9
10					0.997816	10
$X = 81$						
6	0.999998	0.000004	0.000000	0.000000	0.000000	6
7		0.999974	0.000020	0.000000	0.000000	7
8			0.999839	0.000072	0.000000	8
9				0.999353	0.000197	9
10					0.998034	10
$X = 82$						
6	0.999998	0.000003	0.000000	0.000000	0.000000	6
7		0.999977	0.000018	0.000000	0.000000	7
8			0.999859	0.000064	0.000000	8
9				0.999425	0.000177	9
10					0.998231	10
$X = 83$						
6	0.99999 8	0.000003	0.000000	0.000000	0.000000	6
7		0.999981	0.000015	0.000000	0.000000	7
8			0.999877	0.000057	0.000000	8
9				0.999489	0.000159	9
10					0.998408	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 84$						
6	0.999999	0.000002	0.000000	0.000000	0.000000	6
7		0.999983	0.000013	0.000000	0.000000	7
8			0.999892	0.000050	0.000000	8
9				0.999546	0.000143	9
10					0.998567	10
$X = 85$						
6	0.999999	0.000002	0.000000	0.000000	0.000000	6
7		0.999986	0.000012	0.000000	0.000000	7
8			0.999906	0.000045	0.000000	8
9				0.999596	0.000129	9
10					0.998710	10
$X = 86$						
6	0.999999	0.000002	0.000000	0.000000	0.000000	6
7		0.999988	0.000010	0.000000	0.000000	7
8			0.999918	0.000040	0.000000	8
9				0.999641	0.000116	9
10					0.998839	10
$X = 87$						
6	0.999999	0.000001	0.000000	0.000000	0.000000	6
7		0.999990	0.000009	0.000000	0.000000	7
8			0.999928	0.000035	0.000000	8
9				0.999681	0.000104	9
10					0.998955	10
$X = 88$						
6	0.999999	0.000001	0.000000	0.000000	0.000000	6
7		0.999991	0.000008	0.000000	0.000000	7
8			0.999937	0.000032	0.000000	8
9				0.999716	0.000094	9
10					0.999060	10
$X = 89$						
6	0.999999	0.000001	0.000000	0.000000	0.000000	6
7		0.999992	0.000007	0.000000	0.000000	7
8			0.999945	0.000028	0.000000	8
9				0.999748	0.000085	9
10					0.999154	10
$X = 90$						
6	1.000000	0.000001	0.000000	0.000000	0.000000	6
7		0.999993	0.000006	0.000000	0.000000	7
8			0.999952	0.000025	0.000000	8
9				0.999776	0.000076	9
10					0.999238	10
$X = 91$						
6	1.000000	0.000001	0.000000	0.000000	0.000000	6
7		0.999994	0.000005	0.000000	0.000000	7
8			0.999958	0.000022	0.000000	8
9				0.999801	0.000069	9
10					0.999314	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 92$						
6	1.000000	0.000001	0.000000	0.000000	0.000000	6
7		0.999995	0.000005	0.000000	0.000000	7
8			0.999963	0.000020	0.000000	8
9				0.999823	0.000062	9
10					0.999383	10
$X = 93$						
6	1.000000	0.000001	0.000000	0.000000	0.000000	6
7		0.999996	0.000004	0.000000	0.000000	7
8			0.999968	0.000017	0.000000	8
9				0.999843	0.000056	9
10					0.999445	10
$X = 94$						
6	1.000000	0.000001	0.000000	0.000000	0.000000	6
7		0.999996	0.000004	0.000000	0.000000	7
8			0.999972	0.000016	0.000000	8
9				0.999860	0.000050	9
10					0.999500	10
$X = 95$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999997	0.000003	0.000000	0.000000	7
8			0.999975	0.000014	0.000000	8
9				0.999876	0.000045	9
10					0.999550	10
$X = 96$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999997	0.000003	0.000000	0.000000	7
8			0.999978	0.000012	0.000000	8
9				0.999889	0.000040	9
10					0.999595	10
$X = 97$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999998	0.000002	0.000000	0.000000	7
8			0.999981	0.000011	0.000000	8
9				0.999902	0.000036	9
10					0.999636	10
$X = 98$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999998	0.000002	0.000000	0.000000	7
8			0.999983	0.000010	0.000000	8
9				0.999913	0.000033	9
10					0.999672	10
$X = 99$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999998	0.000002	0.000000	0.000000	7
8			0.999985	0.000009	0.000000	8
9				0.999922	0.000030	9
10					0.999705	10

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 6$	$R = 7$	$R = 8$	$R = 9$	$R = 10$	K
$X = 100$						
6	1.000000	0.000000	0.000000	0.000000	0.000000	6
7		0.999999	0.000002	0.000000	0.000000	7
8			0.999987	0.000008	0.000000	8
9				0.999931	0.000027	9
10					0.999734	10
K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 1$						
1	0.090909	0.083333	0.076923	0.071429	0.066667	1
$X = 2$						
1	0.008264	0.006944	0.005917	0.005102	0.004444	1
2	0.016529	0.013889	0.011834	0.010204	0.008889	2
$X = 3$						
1	0.000751	0.000579	0.000455	0.000364	0.000296	1
2	0.004508	0.003472	0.002731	0.002187	0.001778	2
3	0.004508	0.003472	0.002731	0.002187	0.001778	3
$X = 4$						
1	0.000068	0.000048	0.000035	0.000026	0.000020	1
2	0.000956	0.000675	0.000490	0.000364	0.000277	2
3	0.002459	0.001736	0.001260	0.000937	0.000711	3
4	0.001639	0.001157	0.000840	0.000625	0.000474	4
$X = 5$						
2	0.000186	0.000121	0.000081	0.000056	0.000040	2
3	0.000931	0.000603	0.000404	0.000279	0.000198	3
4	0.001490	0.000965	0.000646	0.000446	0.000316	4
5	0.000745	0.000482	0.000323	0.000223	0.000158	5
$X = 6$						
3	0.000305	0.000181	0.000112	0.000072	0.000047	3
4	0.000881	0.000522	0.000323	0.000207	0.000137	4
5	0.001016	0.000603	0.000373	0.000239	0.000158	5
6	0.000406	0.000241	0.000149	0.000096	0.000063	6
$X = 7$						
4	0.000431	0.000234	0.000134	0.000080	0.000049	4
5	0.000862	0.000469	0.000268	0.000159	0.000098	5
6	0.000776	0.000422	0.000241	0.000143	0.000088	6
7	0.000259	0.000141	0.000080	0.000048	0.000029	7
$X = 8$						
4	0.000190	0.000095	0.000050	0.000028	0.000016	4
5	0.000588	0.000293	0.000154	0.000085	0.000049	5
6	0.000893	0.000445	0.000235	0.000130	0.000075	6
7	0.000658	0.000328	0.000173	0.000096	0.000055	7
8	0.000188	0.000094	0.000049	0.000027	0.000016	8
$X = 9$						
5	0.000354	0.000162	0.000079	0.000040	0.000022	5
6	0.000808	0.000369	0.000180	0.000092	0.000050	6
7	0.000988	0.000451	0.000220	0.000113	0.000061	7
8	0.000616	0.000281	0.000137	0.000070	0.000038	8
9	0.000154	0.000070	0.000034	0.000018	0.000009	9

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 10$						
6	0.000634	0.000265	0.000119	0.000057	0.000029	6
7	0.001143	0.000479	0.000215	0.000102	0.000051	7
8	0.001166	0.000488	0.000219	0.000105	0.000052	8
9	0.000630	0.000264	0.000118	0.000056	0.000028	9
10	0.000140	0.000059	0.000026	0.000013	0.000006	10
$X = 11$						
6	0.000453	0.000174	0.000072	0.000032	0.000015	6
7	0.001130	0.000434	0.000180	0.000080	0.000037	7
8	0.001679	0.000645	0.000267	0.000118	0.000055	8
9	0.001469	0.000564	0.000234	0.000103	0.000048	9
10	0.000700	0.000269	0.000111	0.000049	0.000023	10
11	0.000140	0.000054	0.000022	0.000010	0.000005	11
$X = 12$						
6	0.000304	0.000107	0.000041	0.000017	0.000007	6
7	0.001008	0.000355	0.000136	0.000056	0.000024	7
8	0.002043	0.000719	0.000275	0.000113	0.000049	8
9	0.002576	0.000907	0.000347	0.000143	0.000062	9
10	0.001971	0.000694	0.000266	0.000109	0.000048	10
11	0.000839	0.000295	0.000113	0.000046	0.000020	11
12		0.000054	0.000021	0.000008	0.000004	12
$X = 13$						
7	0.000834	0.000269	0.000095	0.000036	0.000015	7
8	0.002219	0.000716	0.000253	0.000096	0.000039	8
9	0.003779	0.001219	0.000431	0.000164	0.000067	9
10	0.004134	0.001334	0.000471	0.000180	0.000073	10
11	0.002811	0.000907	0.000320	0.000122	0.000050	11
12		0.000349	0.000123	0.000047	0.000019	12
13			0.000021	0.000008	0.000003	13
$X = 14$						
7	0.000655	0.000194	0.000063	0.000022	0.000009	7
8	0.002220	0.000657	0.000214	0.000076	0.000029	8
9	0.004907	0.001451	0.000473	0.000168	0.000064	9
10	0.007193	0.002128	0.000694	0.000246	0.000094	10
11	0.006944	0.002054	0.000670	0.000237	0.000090	11
12		0.001256	0.000410	0.000145	0.000055	12
13			0.000144	0.000051	0.000019	13
14				0.000008	0.000003	14
$X = 15$						
7	0.000493	0.000134	0.000040	0.000013	0.000005	7
8	0.002091	0.000567	0.000171	0.000056	0.000020	8
9	0.005831	0.001581	0.000476	0.000157	0.000056	9
10	0.011000	0.002982	0.000898	0.000295	0.000105	10
11	0.014138	0.003833	0.001154	0.000380	0.000135	11
12		0.003310	0.000996	0.000328	0.000116	12
13			0.000554	0.000182	0.000065	13
14				0.000059	0.000021	14
15					0.000003	15
$X = 16$						
7	0.000360	0.000089	0.000025	0.000008	0.000003	7

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
8	0.001879	0.000467	0.000130	0.000040	0.000013	8
9	0.006482	0.001611	0.000448	0.000137	0.000045	9
10	0.015301	0.003803	0.001057	0.000323	0.000107	10
11	0.025138	0.006248	0.001736	0.000530	0.000176	11
12		0.007143	0.001985	0.000606	0.000201	12
13			0.001550	0.000474	0.000157	13
14				0.000241	0.000080	14
15					0.000024	15
$X = 17$						
8	0.001629	0.000371	0.000095	0.000027	0.000008	8
9	0.006841	0.001559	0.000400	0.000113	0.000035	9
10	0.019803	0.004512	0.001157	0.000328	0.000102	10
11	0.040439	0.009213	0.002363	0.000670	0.000207	11
12		0.013391	0.003434	0.000974	0.000302	12
13			0.003535	0.001003	0.000310	13
14				0.000714	0.000221	14
15					0.000104	15
$X = 18$						
8	0.001371	0.000286	0.000068	0.000018	0.000005	8
9	0.006930	0.001447	0.000343	0.000090	0.000026	9
10	0.024222	0.005058	0.001198	0.000315	0.000091	10
11	0.060242	0.012581	0.002978	0.000785	0.000227	11
12		0.022604	0.005351	0.001410	0.000407	12
13			0.006969	0.001836	0.000530	13
14				0.001717	0.000496	14
15					0.000325	15
$X = 19$						
8	0.001127	0.000216	0.000047	0.000012	0.000003	8
9	0.006791	0.001300	0.000284	0.000070	0.000019	9
10	0.028320	0.005421	0.001185	0.000290	0.000078	10
11	0.084464	0.016169	0.003534	0.000864	0.000233	11
12		0.035184	0.007689	0.001881	0.000507	12
13			0.012320	0.003014	0.000812	13
14				0.003553	0.000958	14
15					0.000821	15
$X = 20$						
8	0.000909	0.000160	0.000032	0.000007	0.000002	8
9	0.006479	0.001137	0.000229	0.000052	0.000013	9
10	0.031919	0.005601	0.001130	0.000257	0.000065	10
11	0.112784	0.019791	0.003992	0.000907	0.000228	11
12		0.051354	0.010359	0.002353	0.000592	12
13			0.020009	0.004545	0.001144	13
14				0.006567	0.001652	14
15					0.001779	15
$X = 21$						
8	0.000722	0.000116	0.000022	0.000005	0.000001	8
9	0.006045	0.000972	0.000181	0.000038	0.000009	9
10	0.034907	0.005615	0.001046	0.000221	0.000052	10
11	0.144704	0.023277	0.004334	0.000914	0.000215	11
12		0.071145	0.013248	0.002794	0.000656	12
13			0.030369	0.006405	0.001504	13

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
14				0.011112	0.002610	14
15					0.003431	15
$X = 22$						
8	0.000566	0.000083	0.000014	0.000003	0.000001	8
9	0.005536	0.000816	0.000140	0.000027	0.000006	9
10	0.037229	0.005490	0.000944	0.000185	0.000041	10
11	0.179611	0.026484	0.004552	0.000892	0.000195	11
12		0.094422	0.016230	0.003179	0.000697	12
13			0.043616	0.008543	0.001872	13
14				0.017517	0.003840	14
15					0.006041	15
$X = 23$						
8	0.000439	0.000059	0.000009	0.000002	0.000000	8
9	0.004993	0.000675	0.000107	0.000019	0.000004	9
10	0.038878	0.005255	0.000834	0.000152	0.000031	10
11	0.216840	0.029309	0.004650	0.000846	0.000173	11
12		0.120906	0.019183	0.003489	0.000714	12
13			0.059846	0.010884	0.002227	13
14				0.026060	0.005331	14
15					0.009880	15
$X = 24$						
9	0.004444	0.000551	0.000081	0.000014	0.000003	9
10	0.039883	0.004941	0.000724	0.000122	0.000023	10
11	0.255718	0.031684	0.004640	0.000784	0.000150	11
12		0.150215	0.022000	0.003715	0.000709	12
13			0.079029	0.013346	0.002548	13
14				0.036944	0.007054	14
15					0.015211	15
$X = 25$						
9	0.003912	0.000444	0.000060	0.000009	0.000002	9
10	0.040297	0.004577	0.000619	0.000097	0.000017	10
11	0.295601	0.033573	0.004539	0.000712	0.000127	11
12		0.181899	0.024591	0.003856	0.000687	12
13			0.101029	0.015843	0.002823	13
14				0.050290	0.008962	14
15					0.022265	15
$X = 26$						
9	0.003411	0.000355	0.000044	0.000006	0.000001	9
10	0.040189	0.004184	0.000522	0.000076	0.000013	10
11	0.335898	0.034971	0.004364	0.000635	0.000106	11
12		0.215472	0.026889	0.003915	0.000651	12
13			0.125620	0.018292	0.003042	13
14				0.066133	0.010999	14
15					0.031227	15
$X = 27$						
9	0.002950	0.000281	0.000032	0.000004	0.000001	9
10	0.039636	0.003783	0.000436	0.000059	0.000009	10
11	0.376087	0.035892	0.004134	0.000559	0.000087	11
12		0.250442	0.028849	0.003901	0.000606	12
13			0.152509	0.020621	0.003201	13

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
14				0.084424	0.013106	14
15					0.042226	15
$X = 28$						
9	0.002533	0.000222	0.000024	0.000003	0.000000	9
10	0.038715	0.003387	0.000360	0.000045	0.000007	10
11	0.415724	0.036368	0.003867	0.000486	0.000070	11
12		0.286334	0.030446	0.003823	0.000554	12
13			0.181358	0.022770	0.003299	13
14				0.105045	0.015220	14
15					0.055332	15
$X = 29$						
9	0.002162	0.000173	0.000017	0.000002	0.000000	9
10	0.037498	0.003007	0.000295	0.000034	0.000005	10
11	0.454438	0.036442	0.003577	0.000417	0.000056	11
12		0.322703	0.031674	0.003693	0.000499	12
13			0.211804	0.024693	0.003339	13
14				0.127815	0.017284	14
15					0.070552	15
$X = 30$						
9	0.001836	0.000135	0.000012	0.000001	0.000000	9
10	0.036055	0.002650	0.000240	0.000026	0.000003	10
11	0.491936	0.036162	0.003276	0.000355	0.000045	11
12		0.359145	0.032539	0.003523	0.000445	12
13			0.243478	0.026358	0.003327	13
14				0.152509	0.019249	14
15					0.087836	15
$X = 31$						
9	0.001552	0.000105	0.000009	0.000001	0.000000	9
10	0.034447	0.002321	0.000194	0.000020	0.000002	10
11	0.527991	0.035578	0.002975	0.000299	0.000035	11
12		0.395307	0.033060	0.003323	0.000391	12
13			0.276017	0.027747	0.003269	13
14				0.178867	0.021070	14
15					0.107085	15
$X = 32$						
9	0.001307	0.000081	0.000006	0.000001	0.000000	9
10	0.032726	0.002021	0.000156	0.000015	0.000002	10
11	0.562438	0.034741	0.002682	0.000250	0.000028	11
12		0.430885	0.033264	0.003105	0.000341	12
13			0.309078	0.028851	0.003172	13
14				0.206614	0.022716	14
15					0.128155	15
$X = 33$						
9	0.001097	0.000062	0.000004	0.000000	0.000000	9
10	0.030939	0.001752	0.000125	0.000011	0.000001	10
11	0.595164	0.033699	0.002401	0.000208	0.000021	11
12		0.465626	0.033181	0.002876	0.000295	12
13			0.342342	0.029673	0.003045	13
14				0.235465	0.024162	14
15					0.150871	15

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 34$						
9	0.000918	0.000048	0.000003	0.000000	0.000000	9
10	0.029124	0.001512	0.000099	0.000008	0.000001	10
11	0.626103	0.032496	0.002138	0.000172	0.000016	11
12		0.499325	0.032845	0.002644	0.000253	12
13			0.375523	0.030224	0.002895	13
14				0.265138	0.025393	14
15					0.175034	15
$X = 35$						
9	0.000766	0.000036	0.000002	0.000000	0.000000	9
10	0.027310	0.001299	0.000079	0.000006	0.000001	10
11	0.655227	0.031174	0.001893	0.000141	0.000013	11
12		0.531821	0.032292	0.002413	0.000216	12
13			0.408368	0.030520	0.002728	13
14				0.295362	0.026402	14
15					0.200427	15
$X = 36$						
9	0.000637	0.000028	0.000002	0.000000	0.000000	9
10	0.025524	0.001113	0.000062	0.000004	0.000000	10
11	0.682538	0.029767	0.001668	0.000116	0.000010	11
12		0.562995	0.031555	0.002190	0.000183	12
13			0.440660	0.030581	0.002551	13
14				0.325882	0.027188	14
15					0.226829	15
$X = 37$						
9	0.000530	0.000021	0.000001	0.000000	0.000000	9
10	0.023783	0.000951	0.000049	0.000003	0.000000	10
11	0.708061	0.028307	0.001465	0.000094	0.000007	11
12		0.592762	0.030668	0.001976	0.000154	12
13			0.472215	0.030430	0.002370	13
14				0.356464	0.027757	14
15					0.254018	15
$X = 38$						
9	0.000439	0.000016	0.000001	0.000000	0.000000	9
10	0.022102	0.000810	0.000039	0.000002	0.000000	10
11	0.731844	0.026820	0.001281	0.000077	0.000006	11
12		0.621069	0.029661	0.001775	0.000129	12
13			0.502882	0.030092	0.002187	13
14				0.386894	0.028118	14
15					0.281775	15
$X = 39$						
10	0.020492	0.000688	0.000030	0.000002	0.000000	10
11	0.753946	0.025327	0.001117	0.000062	0.000004	11
12		0.647889	0.028561	0.001578	0.000108	12
13			0.532543	0.029590	0.002007	13
14				0.416986	0.028285	14
15					0.309893	15
$X = 40$						
10	0.018960	0.000584	0.000024	0.000001	0.000000	10
11	0.774438	0.023848	0.000970	0.000050	0.000003	11

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC:0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
12		0.673216	0.027395	0.001413	0.000089	12
13			0.561104	0.028951	0.001833	13
14				0.446576	0.028273	14
15					0.338178	15
$X = 41$						
10	0.017509	0.000494	0.000019	0.000001	0.000000	10
11	0.793397	0.022395	0.000841	0.000040	0.000002	11
12		0.697064	0.026183	0.001254	0.000074	12
13			0.588499	0.028195	0.001666	13
14				0.475527	0.028098	14
15					0.366451	15
$X = 42$						
10	0.016143	0.000418	0.000014	0.000001	0.000000	10
11	0.810907	0.020982	0.000728	0.000032	0.000002	11
12		0.719459	0.024946	0.001110	0.000061	12
13			0.614683	0.027346	0.001508	13
14				0.503722	0.027780	14
15					0.394549	15
$X = 43$						
10	0.014862	0.000353	0.000011	0.000000	0.000000	10
11	0.827050	0.019617	0.000628	0.000026	0.000001	11
12		0.740441	0.023698	0.000979	0.000050	12
13			0.639629	0.026423	0.001360	13
14				0.531068	0.027336	14
15					0.422330	15
$X = 44$						
10	0.013664	0.000297	0.000009	0.000000	0.000000	10
11	0.841912	0.018305	0.000541	0.000021	0.000001	11
12		0.760058	0.022455	0.000861	0.000041	12
13			0.663327	0.025445	0.001222	13
14				0.557491	0.026783	14
15					0.449665	15
$X = 45$						
10	0.012549	0.000250	0.000007	0.000000	0.000000	10
11	0.855576	0.017052	0.000465	0.000017	0.000001	11
12		0.778363	0.021227	0.000756	0.000034	12
13			0.685782	0.024427	0.001095	13
14				0.582936	0.026138	14
15					0.476448	15
$X = 46$						
10	0.011512	0.000210	0.000005	0.000000	0.000000	10
11	0.868125	0.015860	0.000399	0.000013	0.000001	11
12		0.795415	0.020023	0.000662	0.000028	12
13			0.707009	0.023385	0.000979	13
14				0.607363	0.025418	14
15					0.502586	15
$X = 47$						
10	0.010551	0.000177	0.000004	0.000000	0.000000	10
11	0.879637	0.014731	0.000342	0.000011	0.000000	11

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
12		0.811275	0.018852	0.000579	0.000023	12
13			0.727032	0.022329	0.000872	13
14				0.630748	0.024637	14
15					0.528004	15
$X = 48$						
10	0.009662	0.000148	0.000003	0.000000	0.000000	10
11	0.890187	0.013666	0.000293	0.000008	0.000000	11
12		0.826006	0.017717	0.000505	0.000018	12
13			0.745884	0.021272	0.000775	13
14				0.653077	0.023808	14
15					0.552641	15
$X = 49$						
10	0.008841	0.000124	0.000002	0.000000	0.000000	10
11	0.899849	0.012663	0.000251	0.000007	0.000000	11
12		0.839672	0.016625	0.000440	0.000015	12
13			0.763602	0.020222	0.000688	13
14				0.674349	0.022945	14
15					0.576450	15
$X = 50$						
10	0.008085	0.000104	0.000002	0.000000	0.000000	10
11	0.908691	0.011722	0.000214	0.000005	0.000000	11
12		0.852335	0.015578	0.000383	0.000012	12
13			0.780227	0.019186	0.000609	13
14				0.694571	0.022057	14
15					0.599395	15
$X \times 51$						
10	0.007389	0.000087	0.000001	0.000000	0.000000	10
11	0.916776	0.010840	0.000183	0.000004	0.000000	11
12		0.864057	0.014577	0.000333	0.000010	12
13			0.795804	0.018171	0.000539	13
14				0.713757	0.021156	14
15					0.621452	15
$X = 52$						
10	0.006749	0.000073	0.000001	0.000000	0.000000	10
11	0.924164	0.010017	0.000156	0.000003	0.000000	11
12		0.874897	0.013625	0.000289	0.000008	12
13			0.810382	0.017182	0.000475	13
14				0.731928	0.020248	14
15					0.642608	15
$X = 53$						
10	0.006162	0.000061	0.000001	0.000000	0.000000	10
11	0.930914	0.009249	0.000133	0.000003	0.000000	11
12		0.884914	0.012721	0.000250	0.000006	12
13			0.824006	0.016223	0.000419	13
14				0.749111	0.019342	14
15					0.662856	15
$X = 54$						
10	0.005623	0.000051	0.000001	0.000000	0.000000	10
11	0.937075	0.008535	0.000113	0.000002	0 000000	11
12		0.894164	0.011865	0.000217	0.000005	12

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
13			0.836727	0.015297	0.000369	13
14				0.765334	0.018443	14
15					0.682197	15
$X = 55$						
10	0.005129	0.000043	0.000001	0.000000	0.000000	10
11	0.942698	0.007870	0.000096	0.000002	0.000000	11
12		0.902698	0.011057	0.000188	0.000004	12
13			0.848592	0.014406	0.000324	13
14				0.780631	0.017558	14
15					0.700641	15
$X = 56$						
10	0.004678	0.000036	0.000000	0.000000	0.000000	10
11	0.947828	0.007254	0.000082	0.000001	0.000000	11
12		0.910569	0.010295	0.000162	0.000003	12
13			0.859649	0.013551	0.000284	13
14				0.795037	0.016690	14
15					0.718198	15
$X = 57$						
10	0.004264	0.000030	0.000000	0.000000	0.000000	10
11	0.952505	0.006682	0.000070	0.000001	0.000000	11
12		0.917823	0.009579	0.000140	0.000003	12
13			0.869944	0.012734	0.000249	13
14				0.808588	0.015843	14
15					0.734888	15
$X = 58$						
10	0.003886	0.000025	0.000000	0.000000	0.000000	10
11	0.956770	0.006153	0.000059	0.000001	0.000000	11
12		0.924505	0.008907	0.000121	0.000002	12
13			0.879523	0.011955	0.000219	13
14				0.821322	0.015019	14
15					0.750730	15
$X = 59$						
10	0.003541	0.000021	0.000000	0.000000	0.000000	10
11	0.960656	0.005663	0.000050	0.000001	0.000000	11
12		0.930658	0.008276	0.000104	0.000002	12
13			0.888429	0.011213	0.000191	13
14				0.833277	0.014222	14
15					0.765750	15
$X = 60$						
10	0.003225	0.000017	0.000000	0.000000	0.000000	10
11	0.964197	0.005210	0.000043	0.000001	0.000000	11
12		0.936321	0.007686	0.000090	0.000001	12
13			0.896705	0.010509	0.000167	13
14				0.844490	0.013452	14
15					0.779972	15
$X = 61$						
10	0.002938	0.000015	0.000000	0.000000	0.000000	10
11	0.967422	0.004792	0.000036	0.000000	0.000000	11
12		0.941531	0.007134	0.000078	0.000001	12

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
13			0.904391	0.009842	0.000146	13
14				0.954999	0.012712	14
15					0.793424	15
$X = 62$						
10	0.002675	0.000012	0.000000	0.000000	0.000000	10
11	0.970360	0.004406	0.000031	0.000000	0.000000	11
12		0.946323	0.006619	0.000067	0.000001	12
13			0.911526	0.009211	0.000128	13
14				0.864842	0.012001	14
15					0.806136	15
$X = 63$						
10	0.002435	0.000010	0.000000	0.000000	0.000000	10
11	0.973035	0.004050	0.000026	0.000000	0.000000	11
12		0.950729	0.006138	0.000058	0.000001	12
13			0.918145	0.008615	0.000112	13
14				0.874053	0.011320	14
15					0.818137	15
$X = 64$						
10	0.002217	0.000008	0.000000	0.000000	0.000000	10
11	0.975470	0.003722	0.000022	0.000000	0.000000	11
12		0.954779	0.005690	0.000050	0.000001	12
13			0.924283	0.008054	0.000097	13
14				0.882668	0.010670	14
15					0.829457	15
$X = 65$						
10	0.002018	0.000007	0.000000	0.000000	0.000000	10
11	0.977687	0.003419	0.000019	0.000000	0.000000	11
12		0.958500	0.005273	0.000043	0.000000	12
13			0.929973	0.007524	0.000085	13
14				0.890722	0.010049	14
15					0.840127	15
$X = 66$						
10	0.001836	0.000006	0.000000	0.000000	0.000000	10
11	0.979705	0.003141	0.000016	0.000000	0.000000	11
12		0.961920	0.004885	0.000037	0.000000	12
13			0.935246	0.007026	0.000074	13
14				0.898246	0.009459	14
15					0.850176	15
$X = 67$						
10	0.001671	0.000005	0.000000	0.000000	0.000000	10
11	0.981541	0.002885	0.000014	0.000000	0.000000	11
12		0.965061	0.004524	0.000032	0.000000	12
13			0.940131	0.006559	0.000064	13
14				0.905272	0.008897	14
15					0.859634	15
$X = 68$						
10	0.001520	0.000004	0.000000	0.000000	0.000000	10
11	0.983212	0.002649	0.000011	0.000000	0.000000	11
12		0.967945	0.004188	0.000027	0.000000	12

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
13			0.944654	0.006119	0.000056	13
14				0.911831	0.008364	14
15					0.868531	15
$X = 69$						
10	0.001383	0.000003	0.000000	0.000000	0.000000	10
11	0.984732	0.002432	0.000010	0.000000	0.000000	11
12		0.970594	0.003877	0.000023	0.000000	12
13			0.948843	0.005708	0.000049	13
14				0.917951	0.007859	14
15					0.876895	15
$X = 70$						
10	0.001258	0.000003	0.000000	0.000000	0.000000	10
11	0.986115	0.002232	0.000008	0.000000	0.000000	11
12		0.973025	0.003587	0.000020	0.000000	12
13			0.952719	0.005322	0.000043	13
14				0.923658	0.007381	14
15					0.884754	15
$X = 71$						
10	0.001145	0.000002	0.000000	0.000000	0.000000	10
11	0.987373	0.002049	0.000007	0.000000	0.000000	11
12		0.975258	0.003319	0.000017	0.000000	12
13			0.956307	0.004960	0.000037	13
14				0.928980	0.006928	14
15					0.892135	15
$X = 72$						
10	0.001041	0.000002	0.000000	0.000000	0.000000	10
11	0.988518	0.001880	0.000006	0.000000	0.000000	11
12		0.977306	0.003070	0.000015	0.000000	12
13			0.959626	0.004622	0.000032	13
14				0.933940	0.006501	14
15					0.899063	15
$X = 73$						
10	0.000947	0.000002	0.000000	0.000000	0.000000	10
11	0.989559	0.001725	0.000005	0.000000	0.000000	11
12		0.979187	0.002839	0.000013	0.000000	12
13			0.962696	0.004305	0.000028	13
14				0.938561	0.006097	14
15					0.905564	15
$X = 74$						
10	0.000861	0.000001	0.000000	0.000000	0.000000	10
11	0.990506	0.001583	0.000004	0.000000	0.000000	11
12		0.980912	0.002626	0.000011	0.000000	12
13			0.965535	0.004010	0.000024	13
14				0.942867	0.005717	14
15					0.911661	15
$X = 75$						
10	0.000783	0.000001	0.000000	0.000000	0.000000	10
11	0.991367	0.001452	0.000004	0.000000	0.000000	11
12		0.982495	0.002428	0.000009	0.000000	12

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
13			0.968161	0.003733	0.000021	13
14				0.946876	0.005359	14
15					0.917378	15
$X = 76$						
10	0.000712	0.000001	0.000000	0.000000	0.000000	10
11	0.992151	0.001332	0.000003	0.000000	0.000000	11
12		0.983947	0.002244	0.000008	0.000000	12
13			0.970588	0.003475	0.000018	13
14				0.950610	0.005021	14
15					0.922737	15
$X = 77$						
10	0.000648	0.000001	0.000000	0.000000	0.000000	10
11	0.992863	0.001222	0.000003	0.000000	0.000000	11
12		0.985280	0.002074	0.000007	0.000000	12
13			0.972833	0.003235	0.000016	13
14				0.954085	0.004703	14
15					0.927758	15
$X = 78$						
10	0.000589	0.000001	0.000000	0.000000	0.000000	10
11	0.993511	0.001121	0.000002	0.000000	0.000000	11
12		0.986502	0.001917	0.000006	0.000000	12
13			0.974907	0.003010	0.000014	13
14				0.957320	0.004405	14
15					0.932461	15
$X = 79$						
10	0.000536	0.000001	0.000000	0.000000	0.000000	10
11	0.994100	0.001028	0.000002	0.000000	0.000000	11
12		0.987623	0.001772	0.000005	0.000000	12
13			0.976824	0.002800	0.000012	13
14				0.960330	0.004124	14
15					0.936866	15
$X = 80$						
10	0.000487	0.000000	0.000000	0.000000	0.000000	10
11	0.994636	0.000943	0.000002	0.000000	0.000000	11
12		0.988652	0.001637	0.000004	0.000000	12
13			0.978596	0.002605	0.000010	13
14				0.963130	0.003860	14
15					0.940990	15
$X = 81$						
11	0.995123	0.000865	0.000001	0.000000	0.000000	11
12		0.989595	0.001513	0.000004	0.000000	12
13			0.980233	0.002423	0.000009	13
14				0.965735	0.003613	14
15					0.944851	15
$X = 82$						
11	0.995566	0.000793	0.000001	0.000000	0.000000	11
12		0.990460	0.001397	0.000003	0.000000	12
13			0.981745	0.002253	0.000008	13
14				0.968158	0.003380	14
15					0.948463	15

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 83$						
11	0.995969	0.000727	0.000001	0.000000	0.000000	11
12		0.991253	0.001291	0.000003	0.000000	12
13			0.983143	0.002096	0.000007	13
14				0.970412	0.003162	14
15					0.951844	15
$X = 84$						
11	0.996335	0.000667	0.000001	0.000000	0.000000	11
12		0.991981	0.001193	0.000002	0.000000	12
13			0.984434	0.001948	0.000006	13
14				0.972507	0.002958	14
15					0.955006	15
$X = 85$						
11	0.996668	0.000612	0.000001	0.000000	0.000000	11
12		0.992648	0.001102	0.000002	0.000000	12
13			0.985627	0.001811	0.000005	13
14				0.974456	0.002766	14
15					0.957964	15
$X = 86$						
11	0.996971	0.000561	0.000001	0.000000	0.000000	11
12		0.993259	0.001017	0.000002	0.000000	12
13			0.986728	0.001684	0.000004	13
14				0.976267	0.002587	14
15					0.960730	15
$X = 87$						
11	0.997246	0.000514	0.000000	0.000000	0.000000	11
12		0.993820	0.000940	0.000001	0.000000	12
13			0.987746	0.001565	0.000004	13
14				0.977951	0.002418	14
15					0.963317	15
$X = 88$						
11	0.997496	0.000472	0.000000	0.000000	0.000000	11
12		0.994334	0.000868	0.000001	0.000000	12
13			0.988685	0.001455	0.000003	13
14				0.979516	0.002261	14
15					0.965735	15
$X = 89$						
11	0.997724	0.000432	0.000000	0.000000	0.000000	11
12		0.994806	0.000802	0.000001	0.000000	12
13			0.989553	0.001352	0.000003	13
14				0.980971	0.002113	14
15					0.967996	15
$X = 90$						
11	0.997930	0.000396	0.000000	0.000000	0.000000	11
12		0.995238	0.000740	0.000001	0.000000	12
13			0.990355	0.001257	0.000003	13
14				0.982323	0.001975	14
15					0.970109	15

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 91$						
11	0.998119	0.000363	0.000000	0.000000	0.000000	11
12		0.995635	0.000684	0.000001	0.000000	12
13			0.991095	0.001168	0.000002	13
14				0.983580	0.001846	14
15					0.972084	15
$X = 92$						
11	0.998289	0.000333	0.000000	0.000000	0.000000	11
12		0.995998	0.000631	0.000001	0.000000	12
13			0.991778	0.001085	0.000002	13
14				0.984747	0.001725	14
15					0.973930	15
$X = 93$						
11	0.998445	0.000305	0.000000	0.000000	0.000000	11
12		0.996331	0.000583	0.000001	0.000000	12
13			0.992410	0.001008	0.000002	13
14				0.985832	0.001611	14
15					0.975655	15
$X = 94$						
11	0.998586	0.000280	0.000000	0.000000	0.000000	11
12		0.996637	0.000538	0.000001	0.000000	12
13			0.992992	0.000937	0.000001	13
14				0.986840	0.001506	14
15					0.977266	15
$X = 95$						
11	0.998715	0.000257	0.000000	0.000000	0.000000	11
12		0.996917	0.000497	0.000000	0.000000	12
13			0.993531	0.000870	0.000001	13
14				0.987777	0.001407	14
15					0.978772	15
$X = 96$						
11	0.998832	0.000235	0.000000	0.000000	0.000000	11
12		0.997174	0.000459	0.000000	0.000000	12
13			0.994027	0.000808	0.000001	13
14				0.988647	0.001314	14
15					0.980178	15
$X = 97$						
11	0.998938	0.000216	0.000000	0.000000	0.000000	11
12		0.997409	0.000424	0.000000	0.000000	12
13			0.994486	0.000751	0.000001	13
14				0.989456	0.001227	14
15					0.981492	15
$X = 98$						
11	0.999034	0.000198	0.000000	0.000000	0.000000	11
12		0.997625	0.000391	0.000000	0.000000	12
13			0.994910	0.000698	0.000001	13
14				0.990207	0.001146	14
15					0.982719	15

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 11$	$R = 12$	$R = 13$	$R = 14$	$R = 15$	K
$X = 99$						
11	0.999122	0.000181	0.000000	0.000000	0.000000	11
12		0.997823	0.000361	0.000000	0.000000	12
13			0.995301	0.000648	0.000001	13
14				0.990905	0.001071	14
15					0.983866	15
$X = 100$						
11	0.999202	0.000166	0.000000	0.000000	0.000000	11
12		0.998004	0.000333	0.000000	0.000000	12
13			0.995662	0.000602	0.000001	13
14				0.991553	0.001000	14
15					0.984937	15

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 1$						
1	0.062500	0.058824	0.055556	0.052632	0.050000	1
$X = 2$						
1	0.003906	0.003460	0.003086	0.002770	0.002500	1
1	0.007813	0.006920	0.006173	0.005540	0.005000	2
$X = 3$						
1	0.000244	0.000204	0.000171	0.000146	0.000125	1
2	0.001465	0.001221	0.001029	0.000875	0.000750	2
3	0.001465	0.001221	0.001029	0.000875	0.000750	3
$X = 4$						
2	0.000214	0.000168	0.000133	0.000107	0.000088	2
3	0.000549	0.000431	0.000343	0.000276	0.000225	3
4	0.000366	0.000287	0.000229	0.000184	0.000150	4
$X = 5$						
3	0.000143	0.000106	0.000079	0.000061	0.000047	3
4	0.000229	0.000169	0.000127	0.000097	0.000075	4
5	0.000114	0.000085	0.000064	0.000048	0.000037	5
$X = 6$						
$X = 7$						
$X = 8$						
$X = 9$						
$X = 10$						
$X = 11$						
$X = 12$						
$X = 13$						
$X = 14$						
$X = 15$						
$X = 1$						

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 17$						
$X = 18$						
$X = 19$						
$X = 20$						
14	0.000454	0.000135	0.000043	0.000015	0.000005	14
15	0.000489	0.000146	0.000046	0.000016	0.000006	15
16	0.000387	0.000115	0.000037	0.000012	0.000004	16
17		0.000065	0.000021	0.000007	0.000003	17
18			0.000008	0.000003	0.000001	18
19				0.000001	0.000000	19
20					0.000000	20
$X = 21$						
14	0.000673	0.000188	0.000057	0.000018	0.000006	14
15	0.000885	0.000248	0.000075	0.000024	0.000008	15
16	0.000876	0.000245	0.000074	0.000024	0.000008	16
17		0.000180	0.000054	0.000017	0.000006	17
18			0.000029	0.000009	0.000003	18
19				0.000003	0.000001	19
20					0.000000	20
$X = 22$						
13	0.000453	0.000119	0.000034	0.000010	0.000003	13
14	0.000928	0.000245	0.000070	0.000021	0.000007	14
15	0.001460	0.000385	0.000109	0.000033	0.000011	15
16	0.001761	0.000464	0.000132	0.000040	0.000013	16
17		0.000425	0.000121	0.000037	0.000012	17
18			0.000083	0.000025	0.000008	18
19				0.000012	0.000004	19
20					0.000001	20
$X = 23$						
13	0.000505	0.000125	0.000034	0.000010	0.000003	13
14	0.001208	0.000300	0.000080	0.000023	0.000007	14
15	0.002239	0.000555	0.000149	0.000043	0.000013	15
16	0.003221	0.000799	0.000215	0.000062	0.000019	16
17		0.000889	0.000239	0.000069	0.000021	17
18			0.000204	0.000059	0.000018	18
19				0.000038	0.000012	19
20					0.000005	20
$X = 24$						
13	0.000541	0.000126	0.000032	0.000009	0.000003	13
14	0.001499	0.000350	0.000089	0.000024	0.000007	14
15	0.003232	0.000754	0.000191	0.000052	0.000015	15
16	0.005460	0.001274	0.000323	0.000088	0.000026	16
17		0.001688	0.000428	0.000117	0.000034	17
18			0.000442	0.000121	0.000035	18
19				0.000096	0.000028	19
20					0.000017	20
$X = 25$						
13	0.000562	0.000124	0.000030	0.000008	0.000002	13
14	0.001785	0.000392	0.000094	0.000024	0.000007	14

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	R = 16	R = 17	R = 18	R = 19	R = 20	K
15	0.004435	0.000974	0.000233	0.000060	0.000017	15
16	0.008693	0.001910	0.000457	0.000118	0.000033	16
17		0.002963	0.000710	0.000184	0.000051	17
18			0.000871	0.000225	0.000063	18
19				0.000217	0.000060	19
20					0.000045	20
$X = 26$						
13	0.000568	0.000117	0.000027	0.000007	0.000002	13
14	0.002054	0.000425	0.000096	0.000024	0.000006	14
15	0.005832	0.001206	0.000273	0.000067	0.000018	15
16	0.013128	0.002714	0.000614	0.000151	0.000040	16
17		0.004872	0.001102	0.000270	0.000071	17
18			0.001580	0.000387	0.000102	18
19				0.000443	0.000117	19
20					0.000105	20
$X = 27$						
13	0.000560	0.000109	0.000023	0.000005	0.000001	13
14	0.002294	0.000446	0.000095	0.000022	0.000006	14
15	0.007393	0.001439	0.000307	0.000071	0.000018	15
16	0.018959	0.003689	0.000788	0.000183	0.000046	16
17		0.007586	0.001621	0.000377	0.000094	17
18			0.002683	0.000623	0.000156	18
19				0.000830	0.000208	19
20					0.000222	20
$X = 28$						
13	0.000541	0.000099	0.000020	0.000004	0.000001	13
14	0.002498	0.000458	0.000092	0.000020	0.000005	14
15	0.009082	0.001663	0.000336	0.000074	0.000018	15
16	0.026352	0.004826	0.000974	0.000214	0.000051	16
17		0.011275	0.002275	0.000501	0.000119	17
18			0.004304	0.000947	0.000225	18
19				0.001453	0.000346	19
20					0.000430	20
$X = 29$						
13	0.000514	0.000089	0.000017	0.000004	0.000001	13
14	0.002660	0.000458	0.000087	0.000018	0.000004	14
15	0.010856	0.001871	0.000357	0.000074	0.000017	15
16	0.035434	0.006108	0.001164	0.000243	0.000055	16
17		0.016102	0.003069	0.000640	0.000145	17
18			0.006579	0.001372	0.000310	18
19				0.002400	0.000542	19
20					0.000775	20
$X = 30$						
13	0.000480	0.000078	0.000014	0.000003	0.000001	13
14	0.002777	0.000450	0.000081	0.000016	0.000003	14
15	0.012671	0.002056	0.000370	0.000073	0.000016	15
16	0.046290	0.007510	0.001352	0.000267	0.000057	16
17		0.022209	0.003998	0.000790	0.000169	17
18			0.009648	0.001905	0.000409	18
19				0.003772	0.000810	19
20					0.001318	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 31$						
13	0.000442	0.000067	0.000011	0.000002	0.000000	13
14	0.002850	0.000435	0.000074	0.000014	0.000003	14
15	0.014482	0.002211	0.000376	0.000070	0.000014	15
16	0.058961	0.009003	0.001531	0.000286	0.000058	16
17		0.029719	0.005052	0.000945	0.000193	17
18			0.013646	0.002553	0.000521	18
19				0.005677	0.001158	19
20					0.002127	20
$X = 32$						
13	0.000402	0.000058	0.000009	0.000002	0.000000	13
14	0.002880	0.000414	0.000066	0.000012	0.000002	14
15	0.016249	0.002335	0.000375	0.000066	0.000013	15
16	0.073443	0.010554	0.001695	0.000300	0.000058	16
17		0.038722	0.006217	0.001102	0.000213	17
18			0.018698	0.003314	0.000642	18
19				0.008230	0.001594	19
20					0.003285	20
$X = 33$						
14	0.002872	0.000388	0.000059	0.000010	0.000002	14
15	0.017933	0.002426	0.000368	0.000062	0.000011	15
16	0.089692	0.012131	0.001840	0.000309	0.000057	16
17		0.049276	0.007472	0.001255	0.000231	17
18			0.024916	0.004184	0.000770	18
19				0.011545	0.002125	19
20					0.004879	20
$X = 34$						
14	0.002830	0.000360	0.000052	0.000008	0.000001	14
15	0.019505	0.002483	0.000356	0.000057	0.000010	15
16	0.107625	0.013700	0.001962	0.000312	0.000055	16
17		0.061407	0.008795	0.001399	0.000245	17
18			0.032388	0.005153	0.000901	18
19				0.015729	0.002750	19
20					0.007004	20
$X = 35$						
14	0.002758	0.000330	0.000045	0.000007	0.000001	14
15	0.020939	0.002509	0.000339	0.000051	0.000008	15
16	0.127130	0.015231	0.002060	0.000311	0.000052	16
17		0.075107	0.010159	0.001531	0.000254	17
18			0.041183	0.006207	0.001031	18
19				0.020881	0.003468	19
20					0.009753	20
$X = 36$						
14	0.002663	0.000300	0.000038	0.000005	0.000001	14
15	0.022216	0.002505	0.000320	0.000046	0.000007	15
16	0.148069	0.016696	0.002133	0.000305	0.000048	16
17		0.090338	0.011540	0.001648	0.000260	17
18			0.051342	0.007331	0.001157	18
19				0.027088	0.004274	19
20					0.013222	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 37$						
14	0.002549	0.000270	0.000033	0.000004	0.000001	14
15	0.023324	0.002475	0.000299	0.000040	0.000006	15
16	0.170285	0.018072	0.002180	0.000295	0.000044	16
17		0.107035	0.012914	0.001747	0.000262	17
18			0.062882	0.008506	0.001275	18
19				0.034419	0.005159	19
20					0.017496	20
$X = 38$						
14	0.002420	0.000242	0.000028	0.000004	0.000001	14
15	0.024256	0.002423	0.000276	0.000035	0.000005	15
16	0.193609	0.019339	0.002204	0.000282	0.000040	16
17		0.125107	0.014256	0.001827	0.000260	17
18			0.075796	0.009713	0.001383	18
19				0.042925	0.006112	19
20					0.022655	20
$X = 39$						
14	0.002283	0.000215	0.000023	0.000003	0.000000	14
15	0.025009	0.002351	0.000253	0.000031	0.000004	15
16	0.217864	0.020481	0.002204	0.000268	0.000036	16
17		0.144445	0.015545	0.001887	0.000255	17
18			0.090051	0.010933	0.001479	18
19				0.052638	0.007121	19
20					0.028767	20
$X = 40$						
14	0.002139	0.000189	0.000019	0.000002	0.000000	14
15	0.025586	0.002264	0.000230	0.000026	0.000003	15
16	0.242873	0.021489	0.002184	0.000251	0.000032	16
17		0.164927	0.016763	0.001928	0.000248	17
18			0.105596	0.012145	0.001561	18
19				0.063571	0.008170	19
20					0.035888	20
$X = 41$						
14	0.001993	0.000166	0.000016	0.000002	0.000000	14
15	0.025992	0.002164	0.000208	0.000023	0.000003	15
16	0.268459	0.022356	0.002146	0.000234	0.000029	16
17		0.186416	0.017894	0.001950	0.000238	17
18			0.122359	0.013332	0.001628	18
19				0.075716	0.009244	19
20					0.044057	20
$X = 42$						
14	0.001847	0.000145	0.000013	0.000001	0.000000	14
15	0.026236	0.002056	0.000186	0.000019	0.000002	15
16	0.294451	0.023078	0.002092	0.000216	0.000025	16
17		0.208772	0.018927	0.001954	0.000227	17
18			0.140253	0.014478	0.001679	18
19				0.089049	0.010328	19
20					0.053301	20
$X = 43$						
14	0.001704	0.000126	0.000011	0.000001	0.000000	14

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
15	0.026328	0.001942	0.000166	0.000016	0.000002	15
16	0.320686	0.023656	0.002025	0.000198	0.000022	16
17		0.231850	0.019851	0.001941	0.000214	17
18			0.159180	0.015567	0.001715	18
19				0.103527	0.011407	19
20					0.063629	20
$X = 44$						
14	0.001565	0.000109	0.000009	0.000001	0.000000	14
15	0.026280	0.001825	0.000148	0.000014	0.000001	15
16	0.347014	0.024092	0.001948	0.000180	0.000019	16
17		0.255505	0.020662	0.001914	0.000200	17
18			0.179032	0.016587	0.001736	18
19				0.119094	0.012466	19
20					0.075036	20
$X = 45$						
14	0.001432	0.000094	0.000007	0.000001	0.000000	14
15	0.026105	0.001706	0.000130	0.000011	0.000001	15
16	0.373294	0.024392	0.001863	0.000164	0.000016	16
17		0.279597	0.021354	0.001874	0.000186	17
18			0.199693	0.017527	0.001743	18
19				0.135680	0.013492	19
20					0.087502	20
$X = 46$						
14	0.001306	0.000080	0.000006	0.000000	0.000000	14
15	0.025816	0.001588	0.000115	0.000010	0.000001	15
16	0.399399	0.024563	0.001772	0.000147	0.000014	16
17		0.303990	0.021927	0.001823	0.000172	17
18			0.221047	0.018380	0.001736	18
19				0.153208	0.014473	19
20					0.100994	20
$X = 47$						
14	0.001186	0.000069	0.000005	0.000000	0.000000	14
15	0.025427	0.001472	0.000100	0.000008	0.000001	15
16	0.425215	0.024612	0.001677	0.000132	0.000012	16
17		0.328552	0.022382	0.001763	0.000158	17
18			0.242974	0.019140	0.001718	18
19				0.171588	0.015399	19
20					0.115468	20
$X = 48$						
14	0.001075	0.000059	0.000004	0.000000	0.000000	14
15	0.024950	0.001359	0.000087	0.000007	0.000001	15
16	0.450642	0.024549	0.001579	0.000118	0.000010	16
17		0.353164	0.022722	0.001696	0.000145	17
18			0.265356	0.019803	0.001688	18
19				0.190728	0.016261	19
20					0.130867	20
$X = 49$						
14	0.000971	0.000050	0.000003	0.000000	0.000000	14
15	0.024398	0.001251	0.000076	0.000005	0.000000	15
16	0.475592	0.024385	0.001482	0.000105	0.000008	16
17		0.377714	0.022951	0.001623	0.000131	17

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
18			0.288077	0.020367	0.001650	18
19				0.210531	0.017052	19
20					0.147128	20
$X = 50$						
14	0.000875	0.000042	0.000002	0.000000	0.000000	14
15	0.023784	0.001148	0.000066	0.000004	0.000000	15
16	0.499990	0.024128	0.001385	0.000093	0.000007	16
17		0.402098	0.023076	0.001546	0.000119	17
18			0.311029	0.020833	0.001603	18
19				0.230899	0.017767	19
20					0.164180	20
$X = 51$						
14	0.000787	0.000036	0.000002	0.000000	0.000000	14
15	0.023118	0.001050	0.000057	0.000004	0.000000	15
16	0.523773	0.023789	0.001289	0.000082	0.000006	16
17		0.426226	0.023101	0.001466	0.000107	17
18			0.334104	0.021200	0.001550	18
19				0.251731	0.018401	19
20					0.181946	20
$X = 52$						
14	0.000706	0.000030	0.000002	0.000000	0.000000	14
15	0.022411	0.000958	0.000049	0.000003	0.000000	15
16	0.546891	0.023377	0.001197	0.000072	0.000005	16
17		0.450015	0.023036	0.001385	0.000096	17
18			0.357206	0.021473	0.001491	18
19				0.272932	0.018953	19
20					0.200347	20
$X = 53$						
14	0.000632	0.000025	0.000001	0.000000	0.000000	14
15	0.021672	0.000872	0.000042	0.000002	0.000000	15
16	0.569301	0.022904	0.001107	0.000063	0.000004	16
17		0.473392	0.022886	0.001303	0.000086	17
18			0.380241	0.021655	0.001429	18
19				0.294405	0.019422	19
20					0.219300	20
$X = 54$						
14	0.000565	0.000021	0.000001	0.000000	0.000000	14
15	0.020910	0.000792	0.000036	0.000002	0.000000	15
16	0.590973	0.022377	0.001022	0.000055	0.000003	16
17		0.496296	0.022660	0.001223	0.000077	17
18			0.403127	0.021750	0.001363	18
19				0.316061	0.019808	19
20					0.238722	20
$X = 55$						
14	0.000505	0.000018	0.000001	0.000000	0.000000	14
15	0.020133	0.000718	0.000031	0.000002	0.000000	15
16	0.611883	0.021806	0.000940	0.000048	0.000003	16
17		0.518673	0.022366	0.001143	0.000068	17
18			0.425788	0.021764	0.001296	18
19				0.337811	0.020113	19
20					0.258531	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 56$						
14	0.000450	0.000015	0.000001	0.000000	0.000000	14
15	0.019348	0.000649	0.000026	0.000001	0.000000	15
16	0.632017	0.021199	0.000863	0.000042	0.000002	16
17		0.540479	0.022012	0.001066	0.000060	17
18			0.448154	0.021701	0.001227	18
19				0.359574	0.020338	19
20					0.278643	20
$X = 57$						
15	0.018560	0.000586	0.000023	0.000001	0.000000	15
16	0.651365	0.020562	0.000791	0.000036	0.000002	16
17		0.561678	0.021604	0.000991	0.000053	17
18			0.470166	0.021569	0.001159	18
19				0.381276	0.020487	19
20					0.298982	20
$X = 58$						
15	0.017775	0.000528	0.000019	0.000001	0.000000	15
16	0.669925	0.019904	0.000723	0.000031	0.000002	16
17		0.582240	0.021151	0.000919	0.000047	17
18			0.491771	0.021373	0.001091	18
19				0.402845	0.020564	19
20					0.319469	20
$X = 59$						
15	0.016998	0.000475	0.000016	0.000001	0.000000	15
16	0.687700	0.019231	0.000660	0.000027	0.000001	16
17		0.602144	0.020659	0.000851	0.000041	17
18			0.512922	0.021119	0.001024	18
19				0.424217	0.020572	19
20					0.340033	20
$X = 60$						
15	0.016231	0.000427	0.000014	0.000001	0.000000	15
16	0.704698	0.018547	0.000601	0.000023	0.000001	16
17		0.621375	0.020134	0.000785	0.000036	17
18			0.533581	0.020813	0.000959	18
19				0.445336	0.020517	19
20					0.360605	20
$X = 61$						
15	0.015479	0.000383	0.000012	0.000000	0.000000	15
16	0.720929	0.017858	0.000546	0.000020	0.000001	16
17		0.639922	0.019583	0.000724	0.000032	17
18			0.553715	0.020462	0.000896	18
19				0.466149	0.020402	19
20					0.381122	20
$X = 62$						
15	0.014744	0.000344	0.000010	0.000000	0.000000	15
16	0.736409	0.017168	0.000496	0.000017	0.000001	16
17		0.657779	0.019012	0.000666	0.000028	17
18			0.573299	0.020070	0.000834	18
19				0.486611	0.020232	19
20					0.401523	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 63$						
15	0.014029	0.000308	0.000008	0.000000	0.000000	15
16	0.751153	0.016482	0.000450	0.000015	0.000001	16
17		0.674948	0.018424	0.000611	0.000024	17
18			0.592310	0.019644	0.000776	18
19				0.506681	0.020013	19
20					0.421755	20
$X = 64$						
15	0.013334	0.000275	0.000007	0.000000	0.000000	15
16	0.765182	0.015802	0.000407	0.000013	0.000000	16
17		0.691429	0.017825	0.000560	0.000021	17
18			0.610734	0.019189	0.000720	18
19				0.526325	0.019750	19
20					0.441769	20
$X = 65$						
15	0.012661	0.000246	0.000006	0.000000	0.000000	15
16	0.778516	0.015132	0.000368	0.000011	0.000000	16
17		0.707231	0.017220	0.000513	0.000018	17
18			0.628559	0.018710	0.000667	18
19				0.545515	0.019446	19
20					0.461519	20
$X = 66$						
15	0.012012	0.000220	0.000005	0.000000	0.000000	15
16	0.791177	0.014473	0.000333	0.000009	0.000000	16
17		0.722363	0.016611	0.000468	0.000016	17
18			0.645779	0.018211	0.000617	18
19				0.564225	0.019108	19
20					0.480965	20
$X = 67$						
15	0.011386	0.000196	0.000004	0.000000	0.000000	15
16	0.803189	0.013829	0.000300	0.000008	0.000000	16
17		0.736836	0.016003	0.000428	0.000014	17
18			0.662390	0.017696	0.000569	18
19				0.582436	0.018738	19
20					0.500073	20
$X = 68$						
15	0.010785	0.000175	0.000004	0.000000	0.000000	15
16	0.814575	0.013200	0.000271	0.000007	0.000000	16
17		0.750665	0.015397	0.000390	0.000012	17
18			0.678393	0.017170	0.000525	18
19				0.600132	0.018342	19
20					0.518811	20
$X = 69$						
15	0.010209	0.000156	0.000003	0.000000	0.000000	15
16	0.825361	0.012588	0.000244	0.000006	0.000000	16
17		0.763865	0.014798	0.000355	0.000010	17
18			0.693790	0.016635	0.000483	18
19				0.617302	0.017924	19
20					0.537153	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 70$						
15	0.009656	0.000139	0.000003	0.000000	0.000000	15
16	0.835569	0.011994	0.000219	0.000005	0.000000	16
17		0.776452	0.014206	0.000323	0.000009	17
18			0.708588	0.016096	0.000444	18
19				0.633937	0.017486	19
20					0.555077	20
$X = 71$						
15	0.009128	0.000123	0.000002	0.000000	0.000000	15
16	0.845226	0.011419	0.000197	0.000004	0.000000	16
17		0.788446	0.013624	0.000293	0.000008	17
18			0.722793	0.015555	0.000408	18
19				0.650033	0.017034	19
20					0.572563	20
$X = 72$						
15	0.008624	0.000110	0.000002	0.000000	0.000000	15
16	0.854354	0.010863	0.000177	0.000004	0.000000	16
17		0.799865	0.013053	0.000266	0.000007	17
18			0.736417	0.015014	0.000374	18
19				0.665588	0.016569	19
20					0.589597	20
$X = 73$						
15	0.008144	0.000097	0.000002	0.000000	0.000000	15
16	0.862978	0.010327	0.000159	0.000003	0.000000	16
17		0.810728	0.012495	0.000241	0.000006	17
18			0.749470	0.014476	0.000342	18
19				0.680601	0.016096	19
20					0.606167	20
$X = 74$						
15	0.007686	0.000087	0.000001	0.000000	0.000000	15
16	0.871122	0.009812	0.000143	0.000003	0.000000	16
17		0.821055	0.011952	0.000219	0.000005	17
18			0.761966	0.013942	0.000313	18
19				0.695077	0.015616	19
20					0.622262	20
$X = 75$						
15	0.007251	0.000077	0.000001	0.000000	0.000000	15
16	0.878808	0.009316	0.000128	0.000002	0.000000	16
17		0.830867	0.011423	0.000198	0.000004	17
18			0.773918	0.013416	0.000286	18
19				0.709019	0.015133	19
20					0.637879	20
$X = 76$						
15	0.006837	0.000068	0.000001	0.000000	0.000000	15
16	0.886059	0.008840	0.000115	0.000002	0.000000	16
17		0.840183	0.010909	0.000179	0.000004	17
18			0.785340	0.012897	0.000262	18
19				0.722435	0.014648	19
20					0.653012	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 77$						
15	0.006445	0.000061	0.000001	0.000000	0.000000	15
16	0.892896	0.008384	0.000103	0.000002	0.000000	16
17		0.849023	0.010411	0.000162	0.000003	17
18			0.796249	0.012388	0.000239	18
10				0.735332	0.014165	19
20					0.667660	20
$X = 78$						
15	0.006073	0.000054	0.000001	0.000000	0.000000	15
16	0.899341	0.007948	0.000092	0.000001	0.000000	16
17		0.857407	0.009930	0.000146	0.000003	17
18			0.806660	0.011890	0.000218	18
19				0.747720	0.013683	19
20					0.681825	20
$X = 79$						
15	0.005720	0.000048	0.000001	0.000000	0.000000	15
16	0.905414	0.007531	0.000082	0.000001	0.000000	16
17		0.865356	0.009465	0.000132	0.000002	17
18			0.816590	0.011402	0.000198	18
19				0.759610	0.013206	19
20					0.695508	20
$X = 80$						
15	0.005386	0.000042	0.000000	0.000000	0.000000	15
16	0.911134	0.007133	0.000074	0.000001	0.000000	16
17		0.872887	0.009017	0.000119	0.000002	17
18			0.826055	0.010928	0.000180	18
19				0.771012	0.012734	19
20					0.708713	20
$X = 81$						
15	0.005070	0.000037	0.000000	0.000000	0.000000	15
16	0.916521	0.006753	0.000066	0.000001	0.000000	16
17		0.880020	0.008586	0.000108	0.000002	17
18			0.835073	0.010465	0.000164	18
19				0.781940	0.012268	19
20					0.721447	20
$X = 82$						
15	0.004772	0.000033	0.000000	0.000000	0.000000	15
16	0.921591	0.006391	0.000059	0.000001	0.000000	16
17		0.886773	0.008171	0.000097	0.000001	17
18			0.843658	0.010017	0.000149	18
19				0.792405	0.011811	19
20					0.733715	20
$X = 83$						
15	0.004489	0.000029	0.000000	0.000000	0.000000	15
16	0.926363	0.006046	0.000053	0.000001	0.000000	16
17		0.893164	0.007773	0.000087	0.000001	17
18			0.851830	0.009581	0.000136	18
19				0.802422	0.011362	19
20					0.745526	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 84$						
15	0.004223	0.000026	0.000000	0.000000	0.000000	15
16	0.930852	0.005718	0.000047	0.000001	0.000000	16
17		0.899210	0.007391	0.000079	0.000001	17
18			0.859602	0.009160	0.000123	18
19				0.812003	0.010923	19
20					0.756889	20
$X = 85$						
15	0.003971	0.000023	0.000000	0.000000	0.000000	15
16	0.935075	0.005406	0.000042	0.000000	0.000000	16
17		0.904928	0.007024	0.000071	0.000001	17
18			0.866993	0.008752	0.000112	18
19				0.821163	0.010494	19
20					0.767812	20
$X = 86$						
15	0.003734	0.000020	0.000000	0.000000	0.000000	15
16	0.939046	0.005110	0.000037	0.000000	0.000000	16
17		0.910334	0.006674	0.000064	0.000001	17
18			0.874017	0.008359	0.000101	18
19				0.829916	0.010076	19
20					0.778306	20
$X = 87$						
15	0.003510	0.000018	0.000000	0.000000	0.000000	15
16	0.942779	0.004828	0.000033	0.000000	0.000000	16
17		0.915444	0.006338	0.000057	0.000001	17
18			0.880691	0.007979	0.000092	18
19				0.838274	0.009668	19
20					0.788381	20
$X = 88$						
15	0.003299	0.000016	0.000000	0.000000	0.000000	15
16	0.946289	0.004561	0.000030	0.000000	0.000000	16
17		0.920272	0.006018	0.000052	0.000001	17
18			0.887030	0.007614	0.000083	18
19				0.846254	0.009272	19
20					0.798049	20
$X = 89$						
15	0.003100	0.000014	0.000000	0.000000	0.000000	15
16	0.949588	0.004308	0.000027	0.000000	0.000000	16
17		0.924833	0.005712	0.000046	0.000000	17
18			0.893047	0.007262	0.000076	18
19				0.853868	0.008888	19
20					0.807321	20
$X = 90$						
15	0.002912	0.000012	0.000000	0.000000	0.000000	15
16	0.952687	0.004068	0.000024	0.000000	0.000000	16
17		0.929141	0.005420	0.000042	0.000000	17
18			0.898759	0.006924	0.000068	18
19				0.861130	0.008515	19
20					0.816209	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 91$						
15	0.002736	0.000011	0.000000	0.000000	0.000000	15
16	0.955600	0.003840	0.000021	0.000000	0.000000	16
17		0.933209	0.005141	0.000038	0.000000	17
18			0.904179	0.006599	0.000062	18
19				0.868054	0.008155	19
20					0.824724	20
$X = 92$						
15	0.002570	0.000010	0.000000	0.000000	0.000000	15
16	0.958336	0.003625	0.000019	0.000000	0.000000	16
17		0.937049	0.004875	0.000034	0.000000	17
18			0.909320	0.006287	0.000056	18
19				0.874653	0.007806	19
20					0.832879	20
$X = 93$						
15	0.002413	0.000009	0.000000	0.000000	0.000000	15
16	0.960905	0.003421	0.000017	0.000000	0.000000	16
17		0.940674	0.004622	0.000030	0.000000	17
18			0.914195	0.005988	0.000051	18
19				0.880940	0.007469	19
20					0.840685	20
$X = 94$						
15	0.002266	0.000008	0.000000	0.000000	0.000000	15
16	0.963319	0.003227	0.000015	0.000000	0.000000	16
17		0.944094	0.004381	0.000027	0.000000	17
18			0.918817	0.005702	0.000046	18
19				0.886929	0.007144	19
20					0.848154	20
$X = 95$						
15	0.002128	0.000007	0.000000	0.000000	0.000000	15
16	0.965585	0.003045	0.000013	0.000000	0.000000	16
17		0.947322	0.004152	0.000024	0.000000	17
18			0.923198	0.005428	0.000042	18
19				0.892630	0.006830	19
20					0.855297	20
$X = 96$						
15	0.001998	0.000006	0.000000	0.000000	0.000000	15
16	0.967713	0.002872	0.000012	0.000000	0.000000	16
17		0.950366	0.003934	0.000022	0.000000	17
18			0.927350	0.005165	0.000038	18
19				0.898058	0.006528	19
20					0.862127	20
$X = 97$						
15	0.001875	0.000005	0.000000	0.000000	0.000000	15
16	0.969711	0.002709	0.000011	0.000000	0.000000	16
17		0.953238	0.003727	0.000020	0.000000	17
18			0.931284	0.004914	0.000034	18
19				0.903223	0.006237	19
20					0.868655	20

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 16$	$R = 17$	$R = 18$	$R = 19$	$R = 20$	K
$X = 98$						
15	0.001760	0.000005	0.000000	0.000000	0.000000	15
16	0.971586	0.002554	0.000009	0.000000	0.000000	16
17		0.955947	0.003530	0.000018	0.000000	17
18			0.935011	0.004674	0.000031	18
19				0.908137	0.005958	19
20					0.874893	20
$X = 99$						
15	0.001652	0.000004	0.000000	0.000000	0.000000	15
16	0.973347	0.002408	0.000008	0.000000	0.000000	16
17		0.958501	0.003342	0.000016	0.000000	17
18			0.938541	0.004445	0.000028	18
19				0.912811	0.005689	19
20					0.880850	20
$X = 100$						
15	0.001551	0.000004	0.000000	0.000000	0.000000	15
16	0.974999	0.002270	0.000007	0.000000	0.000000	16
17		0.960910	0.003165	0.000014	0.000000	17
18			0.941883	0.004226	0.000025	18
19				0.917256	0.005431	19
20					0.886539	20

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 1$						
1	0.047619	0.045455	0.043478	0.041667	0.040000	1
$X = 2$						
1	0.002268	0.002066	0.001890	0.001736	0.001600	1
2	0.004535	0.004132	0.003781	0.003472	0.003200	2
$X = 3$						
1	0.000108	0.000094	0.000082	0.000072	0.000064	1
2	0.000648	0.000563	0.000493	0.000434	0.000384	2
3	0.000648	0.000563	0.000493	0.000434	0.000384	3
$X = 4$						
2	0.000072	0.000060	0.000050	0.000042	0.000036	2
3	0.000185	0.000154	0.000129	0.000109	0.000092	3
4	0.000123	0.000102	0.000086	0.000072	0.000061	4
$X = 5$						
$X = 6$						
$X = 7$						
$X = 8$						
$X = 9$						
$X = 10$						
$X = 11$						
$X = 12$						
$X = 13$						

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 14$						
$X = 15$						
$X = 16$						
$X = 17$						
$X = 18$						
$X = 19$						
$X = 20$						
$X = 21$						
$X = 22$						
$X = 23$						
$X = 24$						
$X = 25$						
$X = 26$						
$X = 27$						
$X = 28$						
$X = 29$						
$X = 30$						
$X = 31$						
20	0.000469	0.000111	0.000028	0.000007	0.000002	20
21	0.000713	0.000169	0.000042	0.000011	0.000003	21
22		0.000212	0.000053	0.000014	0.000004	22
23			0.000055	0.000015	0.000004	23
24				0.000013	0.000004	24
25					0.000002	25
$X = 32$						
20	0.000689	0.000156	0.000038	0.000010	0.000003	20
21	0.001182	0.000267	0.000064	0.000016	0.000004	21
22		0.000380	0.000092	0.000023	0.000006	22
23			0.000109	0.000028	0.000008	23
24				0.000027	0.000007	24
25					0.000006	25
$X = 33$						
20	0.000975	0.000210	0.000048	0.000012	0.000003	20
21	0.001871	0.000403	0.000093	0.000023	0.000006	21
22		0.000647	0.000149	0.000037	0.000010	22
23			0.000200	0.000049	0.000013	23
24				0.000055	0.000014	24
25					0.000013	25
$X = 34$						
19	0.000523	0.000108	0.000024	0.000006	0.000001	19
20	0.001333	0.000274	0.000060	0.000014	0.000004	20
21	0.002846	0.000585	0.000129	0.000030	0.000008	21
22		0.001050	0.000232	0.000055	0.000014	22
23			0.000350	0.000082	0.000021	23

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
24				0.000104	0.000026	24
25					0.000028	25
$X = 35$						
19	0.000629	0.000123	0.000026	0.000006	0.000001	19
20	0.001768	0.000347	0.000073	0.000017	0.000004	20
21	0.004179	0.000820	0.000173	0.000039	0.000009	21
22		0.001635	0.000345	0.000078	0.000019	22
23			0.000581	0.000131	0.000031	23
24				0.000187	0.000045	24
25					0.000054	25
$X = 36$						
19	0.000738	0.000138	0.000028	0.000006	0.000001	19
20	0.002283	0.000428	0.000086	0.000019	0.000004	20
21	0.005948	0.001114	0.000225	0.000049	0.000011	21
22		0.002456	0.000496	0.000107	0.000025	22
23			0.000927	0.000200	0.000046	23
24				0.000318	0.000073	24
25					0.000098	25
$X = 37$						
19	0.000848	0.000152	0.000029	0.000006	0.000001	19
20	0.002877	0.000515	0.000099	0.000021	0.000005	20
21	0.008230	0.001472	0.000284	0.000059	0.000013	21
22		0.003570	0.000689	0.000143	0.000032	22
23			0.001422	0.000294	0.000065	23
24				0.000518	0.000114	24
25					0.000172	25
$X = 38$						
19	0.000957	0.000163	0.000030	0.000006	0.000001	19
20	0.003548	0.000606	0.000112	0.000022	0.000005	20
21	0.011107	0.001896	0.000350	0.000069	0.000015	21
22		0.005042	0.000931	0.000185	0.000039	22
23			0.002112	0.000419	0.000089	23
24				0.000812	0.000172	24
25					0.000286	25
$X = 39$						
19	0.001062	0.000173	0.000031	0.000006	0.000001	19
20	0.004291	0.000699	0.000124	0.000023	0.000005	20
21	0.014655	0.002388	0.000422	0.000080	0.000016	21
22		0.006938	0.001226	0.000233	0.000047	22
23			0.003043	0.000579	0.000118	23
24				0.001231	0.000251	24
25					0.000458	25
$X = 40$						
19	0.001160	0.000181	0.000030	0.000006	0.000001	19
20	0.005098	0.000793	0.000134	0.000024	0.000005	20
21	0.018946	0.002947	0.000498	0.000091	0.000018	21
22		0.009326	0.001576	0.000287	0.000056	22
23			0.004268	0.000778	0.000152	23
24				0.001810	0.000354	24
25					0.000709	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 41$						
19	0.001251	0.000186	0.000030	0.000005	0.000001	19
20	0.005960	0.000885	0.000143	0.000025	0.000005	20
21	0.024043	0.003570	0.000577	0.000101	0.000019	21
22		0.012273	0.001984	0.000346	0.000065	22
23			0.005844	0.001021	0.000191	23
24				0.002588	0.000485	24
25					0.001062	25
$X = 42$						
19	0.001331	0.000189	0.000029	0.000005	0.000001	19
20	0.006867	0.000973	0.000150	0.000025	0.000005	20
21	0.030003	0.004252	0.000657	0.000110	0.000020	21
22		0.015843	0.002449	0.000410	0.000074	22
23			0.007828	0.001310	0.000236	23
24				0.003608	0.000650	24
25					0.001548	25
$X = 43$						
19	0.001400	0.000189	0.000028	0.000004	0.000001	19
20	0.007808	0.001056	0.000156	0.000025	0.000004	20
21	0.036871	0.004988	0.000738	0.000118	0.000020	21
22		0.020095	0.002972	0.000477	0.000082	22
23			0.010277	0.001648	0.000285	23
24				0.004919	0.000850	24
25					0.002197	25
$X = 44$						
19	0.001457	0.000188	0.000027	0.000004	0.000001	19
20	0.008769	0.001132	0.000160	0.000025	0.000004	20
21	0.044678	0.005770	0.000816	0.000125	0.000021	21
22		0.025083	0.003548	0.000545	0.000090	22
23			0.013248	0.002036	0.000338	23
24				0.006567	0.001090	24
25					0.003048	25
$X = 45$						
19	0.001502	0.000185	0.000025	0.000004	0.000001	19
20	0.009739	0.001200	0.000162	0.000024	0.000004	20
21	0.053447	0.006588	0.000891	0.000131	0.000021	21
22		0.030853	0.004174	0.000615	0.000098	22
23			0.016796	0.002474	0.000394	23
24				0.008603	0.001371	24
25					0.004137	25
$X = 46$						
19	0.001534	0.000181	0.000023	0.000003	0.000001	19
20	0.010705	0.001260	0.000163	0.000023	0.000004	20
21	0.063186	0.007435	0.000962	0.000136	0.000021	21
22		0.037441	0.004845	0.000684	0.000105	22
23			0.020971	0.002960	0.000453	23
24				0.011078	0.001694	24
25					0.005508	25
$X = 47$						
19	0.001555	0.000175	0.000022	0.000003	0.000000	19

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
20	0.011656	0.001309	0.000162	0.000022	0.000003	20
21	0.073891	0.008299	0.001027	0.000139	0.000020	21
22		0.044876	0.005555	0.000752	0.000110	22
23			0.025816	0.003493	0.000513	23
24				0.014038	0.002061	24
25					0.007202	25
$X = 48$						
19	0.001563	0.000168	0.000020	0.000003	0.000000	19
20	0.012582	0.001349	0.000160	0.000021	0.000003	20
21	0.085547	0.009171	0.001086	0.000141	0.000020	21
22		0.053175	0.006296	0.000816	0.000115	22
23			0.031371	0.004067	0.000573	23
24				0.017531	0.002471	24
25					0.009263	25
$X = 49$						
19	0.001561	0.000160	0.000018	0.000002	0.000000	19
20	0.013472	0.001379	0.000156	0.000019	0.000003	20
21	0.098129	0.010042	0.001137	0.000141	0.000019	21
22		0.062346	0.007061	0.000877	0.000119	22
23			0.037667	0.004680	0.000633	23
24				0.021598	0.002922	24
25					0.011734	25
$X = 50$						
19	0.001549	0.000151	0.000016	0.000002	0.000000	19
20	0.014317	0.001399	0.000152	0.000018	0.000002	20
21	0.111601	0.010902	0.001181	0.000141	0.000018	21
22		0.072388	0.007842	0.000934	0.000121	22
23			0.044728	0.005326	0.000692	23
24				0.026278	0.003413	24
25					0.014656	25
$X = 51$						
19	0.001528	0.000142	0.000015	0.000002	0.000000	19
20	0.015111	0.001409	0.000146	0.000017	0.000002	20
21	0.125918	0.011741	0.001217	0.000139	0.000017	21
22		0.083290	0.008631	0.000985	0.000123	22
23			0.052570	0.005999	0.000748	23
24				0.031604	0.003941	24
25					0.018069	25
$X = 52$						
19	0.001499	0.000133	0.000013	0.000001	0.000000	19
20	0.015847	0.001410	0.000140	0.000015	0.000002	20
21	0.141028	0.012552	0.001244	0.000136	0.000016	21
22		0.095031	0.009419	0.001030	0.000123	22
23			0.061200	0.006693	0.000801	23
24				0.037602	0.004501	24
25					0.022010	25
$X = 53$						
19	0.001463	0.000124	0.000012	0.000001	0.000000	19
20	0.016520	0.001404	0.000133	0.000014	0.000002	20
21	0.156875	0.013328	0.001264	0.000132	0.000015	21

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
22		0.107583	0.010200	0.001069	0.000123	22
23			0.070619	0.007401	0.000851	23
24				0.044295	0.005090	24
25					0.026511	25
$X = 54$						
19	0.001421	0.000115	0.000010	0.000001	0.000000	19
20	0.017127	0.001389	0.000126	0.000013	0.000001	20
21	0.173395	0.014062	0.001275	0.000128	0.000014	21
22		0.120911	0.010965	0.001101	0.000121	22
23			0.080819	0.008117	0.000895	23
24				0.051696	0.005703	24
25					0.031601	25
$X = 55$						
19	0.001374	0.000106	0.000009	0.000001	0.000000	19
20	0.017664	0.001367	0.000119	0.000011	0.000001	20
21	0.190522	0.014749	0.001279	0.000123	0.000013	21
22		0.134973	0.011708	0.001127	0.000119	22
23			0.091783	0.008834	0.000936	23
24				0.059812	0.006335	24
25					0.037305	25
$X = 56$						
19	0.001323	0.000098	0.000008	0.000001	0.000000	19
20	0.018132	0.001340	0.000111	0.000010	0.000001	20
21	0.208186	0.015383	0.001276	0.000118	0.000012	21
22		0.149721	0.012422	0.001146	0.000117	22
23			0.103491	0.009546	0.000971	23
24				0.068646	0.006980	24
25					0.043640	25
$X = 57$						
19	0.001270	0.000090	0.000007	0.000001	0.000000	19
20	0.018529	0.001307	0.000104	0.000009	0.000001	20
21	0.226318	0.015963	0.001267	0.000112	0.000011	21
22		0.165105	0.013103	0.001158	0.000113	22
23			0.115913	0.010246	0.001000	23
24				0.078192	0.007632	24
25					0.050619	25
$X = 58$						
19	0.001214	0.000082	0.000006	0.000001	0.000000	19
20	0.018856	0.001270	0.000096	0.000008	0.000001	20
21	0.244847	0.016485	0.001251	0.000106	0.000010	21
22		0.181068	0.013745	0.001164	0.000109	22
23			0.129017	0.010929	0.001024	23
24				0.088438	0.008287	24
25					0.058251	25
$X = 59$						
19	0.001156	0.000074	0.000005	0.000000	0.000000	19
20	0.019114	0.001228	0.000089	0.000007	0.000001	20
21	0.263703	0.016947	0.001231	0.000100	0.000009	21
22		0.197553	0.014345	0.001164	0.000105	22
23			0.142762	0.011589	0.001043	23

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
24				0.099367	0.008939	24
25					0.066538	25
$X = 60$						
19	0.001098	0.000067	0.000005	0.000000	0.000000	19
20	0.019305	0.001184	0.000082	0.000006	0.000001	20
21	0.282817	0.017350	0.001205	0.000094	0.000008	21
22		0.214500	0.014898	0.001159	0.000100	22
23			0.157106	0.012222	0.001056	23
24				0.110956	0.009582	24
25					0.075477	25
$X = 61$						
19	0.001040	0.000061	0.000004	0.000000	0.000000	19
20	0.019432	0.001138	0.000076	0.000006	0.000000	20
21	0.302122	0.017692	0.001175	0.000088	0.000007	21
22		0.231850	0.015403	0.001148	0.000095	22
23			0.172005	0.012824	0.001063	23
24				0.123179	0.010212	24
25					0.085059	25
$X = 62$						
19	0.000982	0.000055	0.000003	0.000000	0.000000	19
20	0.019497	0.001090	0.000069	0.000005	0.000000	20
21	0.321554	0.017974	0.001142	0.000082	0.000006	21
22		0.249542	0.015858	0.001133	0.000090	22
23			0.187407	0.013390	0.001066	23
24				0.136003	0.010824	24
25					0.095271	25
$X = 63$						
19	0.000926	0.000049	0.000003	0.000000	0.000000	19
20	0.019504	0.001041	0.000063	0.000004	0.000000	20
21	0.341051	0.018197	0.001106	0.000076	0.000006	21
22		0.267515	0.016261	0.001113	0.000085	22
23			0.203265	0.013918	0.001063	23
24				0.149393	0.011414	24
25					0.106095	25
$X = 64$						
19	0.000870	0.000044	0.000003	0.000000	0.000000	19
20	0.019457	0.000991	0.000058	0.000004	0.000000	20
21	0.360556	0.018363	0.001068	0.000070	0.000005	21
22		0.285712	0.016612	0.001090	0.000080	22
23			0.219526	0.014405	0.001057	23
24				0.163311	0.011979	24
25					0.117509	25
$X = 65$						
19	0.000816	0.000040	0.000002	0.000000	0.000000	19
20	0.019359	0.000941	0.000052	0.000003	0.000000	20
21	0.380013	0.018474	0.001027	0.000065	0.000005	21
22		0.304075	0.016911	0.001063	0.000075	22
23			0.236137	0.014849	0.001046	23
24				0.177716	0.012514	24
25					0.129488	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 66$						
19	0.000763	0.000035	0.000002	0.000000	0.000000	19
20	0.019214	0.000892	0.000047	0.000003	0.000000	20
21	0.399372	0.018533	0.000986	0.000059	0.000004	21
22		0.322549	0.017158	0.001034	0.000070	22
23			0.253048	0.015250	0.001031	23
24				0.192565	0.013017	24
25					0.142001	25
$X = 67$						
19	0.000713	0.000032	0.000002	0.000000	0.000000	19
20	0.019026	0.000843	0.000043	0.000002	0.000000	20
21	0.418586	0.018542	0.000943	0.000054	0.000004	21
22		0.341082	0.017355	0.001002	0.000065	22
23			0.270206	0.015605	0.001013	23
24				0.207814	0.013486	24
25					0.155018	25
$X = 68$						
19	0.000665	0.000028	0.000001	0.000000	0.000000	19
20	0.018799	0.000795	0.000039	0.000002	0.000000	20
21	0.437612	0.018503	0.000901	0.000050	0.000003	21
22		0.359624	0.017503	0.000969	0.000060	22
23			0.287561	0.015915	0.000992	23
24				0.223420	0.013919	24
25					0.168504	25
$X = 69$						
19	0.000619	0.000025	0.000001	0.000000	0.000000	19
20	0.018537	0.000748	0.000035	0.000002	0.000000	20
21	0.456411	0.018421	0.000858	0.000045	0.000003	21
22		0.378127	0.017603	0.000934	0.000056	22
23			0.305064	0.016181	0.000968	23
24				0.239335	0.014314	24
25					0.182423	25
$X = 70$						
19	0.000575	0.000022	0.000001	0.000000	0.000000	19
20	0.018243	0.000703	0.000031	0.000002	0.000000	20
21	0.474948	0.018298	0.000815	0.000041	0.000002	21
22		0.396548	0.017658	0.000898	0.000052	22
23			0.322667	0.016401	0.000942	23
24				0.255516	0.014670	24
25					0.196737	25
$X = 71$						
19	0.000533	0.000020	0.000001	0.000000	0.000000	19
20	0.017922	0.000659	0.000028	0.000001	0.000000	20
21	0.493192	0.018137	0.000773	0.000038	0.000002	21
22		0.414845	0.017670	0.000861	0.000047	22
23			0.340326	0.016578	0.000914	23
24				0.271917	0.014988	24
25					0.211407	25
$X = 72$						
19	0.000494	0.000017	0.000001	0.000000	0.000000	19

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
20	0.017576	0.000617	0.000025	0.000001	0.000000	20
21	0.511113	0.017942	0.000731	0.000034	0.000002	21
22		0.432982	0.017641	0.000823	0.000044	22
23			0.357996	0.016712	0.000884	23
24				0.288494	0.015265	24
25					0.226395	25
$X = 73$						
19	0.000457	0.000015	0.000001	0.000000	0.000000	19
20	0.017210	0.000577	0.000022	0.000001	0.000000	20
21	0.528690	0.017715	0.000690	0.000031	0.000002	21
22		0.450924	0.017573	0.000786	0.000040	22
23			0.375637	0.016805	0.000854	23
24				0.305206	0.015504	24
25					0.241660	25
$X = 74$						
20	0.016825	0.000538	0.000020	0.000001	0.000000	20
21	0.545899	0.017460	0.000651	0.000028	0.000001	21
22		0.468639	0.017469	0.000749	0.000037	22
23			0.393210	0.016858	0.000822	23
24				0.322011	0.015703	24
25					0.257164	25
$X = 75$						
20	0.016426	0.000502	0.000018	0.000001	0.000000	20
21	0.562725	0.017180	0.000613	0.000025	0.000001	21
22		0.486099	0.017332	0.000712	0.000033	22
23			0.410679	0.016873	0.000790	23
24				0.338869	0.015864	24
25					0.272867	25
$X = 76$						
20	0.016015	0.000467	0.000016	0.000001	0.000000	20
21	0.579151	0.016878	0.000576	0.000023	0.000001	21
22		0.503279	0.017165	0.000676	0.000030	22
23			0.428011	0.016853	0.000757	23
24				0.355742	0.015988	24
25					0.288731	25
$X = 77$						
20	0.015595	0.000434	0.000014	0.000001	0.000000	20
21	0.595167	0.016556	0.000540	0.000020	0.000001	21
22		0.520157	0.016969	0.000640	0.000028	22
23			0.445176	0.016798	0.000725	23
24				0.372595	0.016076	24
25					0.304719	25
$X = 78$						
20	0.015167	0.000403	0.000013	0.000000	0.000000	20
21	0.610762	0.016218	0.000506	0.000018	0.000001	21
22		0.536714	0.016748	0.000606	0.000025	22
23			0.462145	0.016712	0.000692	23
24				0.389393	0.016129	24
25					0.320795	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 79$						
20	0.014735	0.000373	0.000011	0.000000	0.000000	20
21	0.625929	0.015865	0.000474	0.000016	0.000001	21
22		0.552932	0.016504	0.000572	0.000023	22
23			0.478893	0.016596	0.000660	23
24				0.406104	0.016148	24
25					0.336923	25
$X = 80$						
20	0.014300	0.000346	0.000010	0.000000	0.000000	20
21	0.640664	0.015501	0.000443	0.000015	0.000001	21
22		0.568797	0.016239	0.000539	0.000021	22
23			0.495397	0.016452	0.000628	23
24				0.422700	0.016136	24
25					0.353071	25
$X = 81$						
20	0.013864	0.000320	0.000009	0.000000	0.000000	20
21	0.654964	0.015126	0.000413	0.000013	0.000000	21
22		0.584297	0.015957	0.000508	0.000019	22
23			0.511637	0.016284	0.000597	23
24				0.439153	0.016093	24
25					0.369207	25
$X = 82$						
20	0.013428	0.000296	0.000008	0.000000	0.000000	20
21	0.668827	0.014744	0.000385	0.000012	0.000000	21
22		0.599423	0.015658	0.000478	0.000017	22
23			0.527593	0.016092	0.000566	23
24				0.455436	0.016023	24
25					0.385301	25
$X = 83$						
20	0.012994	0.000273	0.000007	0.000000	0.000000	20
21	0.682255	0.014357	0.000359	0.000010	0.000000	21
22		0.614168	0.015346	0.000449	0.000015	22
23			0.543252	0.015879	0.000536	23
24				0.471528	0.015925	24
25					0.401323	25
$X = 84$						
20	0.012564	0.000252	0.000006	0.000000	0.000000	20
21	0.695250	0.013965	0.000334	0.000009	0.000000	21
22		0.628524	0.015022	0.000421	0.000014	22
23			0.558598	0.015647	0.000508	23
24				0.487407	0.015803	24
25					0.417249	25
$X = 85$						
20	0.012139	0.000233	0.000005	0.000000	0.000000	20
21	0.707814	0.013571	0.000310	0.000008	0.000000	21
22		0.642490	0.014688	0.000394	0.000012	22
23			0.573619	0.015399	0.000480	23
24				0.503055	0.015659	24
25					0.433052	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 86$						
20	0.011719	0.000215	0.000005	0.000000	0.000000	20
21	0.719953	0.013177	0.000288	0.000007	0.000000	21
22		0.656061	0.014346	0.000369	0.000011	22
23			0.588307	0.015135	0.000453	23
24				0.518453	0.015493	24
25					0.448711	25
$X = 87$						
20	0.011305	0.000198	0.000004	0.000000	0.000000	20
21	0.731672	0.012782	0.000267	0.000007	0.000000	21
22		0.669238	0.013998	0.000345	0.000010	22
23			0.602654	0.014858	0.000427	23
24				0.533588	0.015307	24
25					0.464203	25
$X = 88$						
20	0.010899	0.000182	0.000004	0.000000	0.000000	20
21	0.742977	0.012390	0.000248	0.000006	0.000000	21
22		0.682020	0.013645	0.000322	0.000009	22
23			0.616652	0.014570	0.000402	23
24				0.548446	0.015104	24
25					0.479510	25
$X = 89$						
20	0.010501	0.000167	0.000003	0.000000	0.000000	20
21	0.753876	0.012000	0.000230	0.000005	0.000000	21
22		0.694410	0.013289	0.000301	0.000008	22
23			0.630298	0.014272	0.000378	23
24				0.563016	0.014886	24
25					0.494615	25
$X = 90$						
20	0.010111	0.000154	0.000003	0.000000	0.000000	20
21	0.764376	0.011614	0.000213	0.000005	0.000000	21
22		0.706410	0.012931	0.000281	0.000007	22
23			0.643587	0.013966	0.000355	23
24				0.577288	0.014653	24
25					0.509501	25
$X = 91$						
20	0.009730	0.000141	0.000002	0.000000	0.000000	20
21	0.774487	0.011233	0.000197	0.000004	0.000000	21
22		0.718025	0.012573	0.000262	0.000006	22
23			0.656518	0.013653	0.000333	23
24				0.591253	0.014407	24
25					0.524153	25
$X = 92$						
20	0.009358	0.000130	0.000002	0.000000	0.000000	20
21	0.784217	0.010857	0.000182	0.000004	0.000000	21
22		0.729258	0.012214	0.000244	0.000006	22
23			0.669091	0.013334	0.000312	23
24				0.604906	0.014151	24
25					0.538561	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 93$						
20	0.008996	0.000119	0.000002	0.000000	0.000000	20
21	0.793575	0.010488	0.000168	0.000003	0.000000	21
22		0.740115	0.011857	0.000227	0.000005	22
23			0.681305	0.013012	0.000293	23
24				0.618240	0.013884	24
25					0.552711	25
$X = 94$						
20	0.008644	0.000109	0.000002	0.000000	0.000000	20
21	0.802571	0.010124	0.000155	0.000003	0.000000	21
22		0.750603	0.011503	0.000211	0.000005	22
23			0.693162	0.012687	0.000274	23
24				0.631252	0.013610	24
25					0.566596	25
$X = 95$						
20	0.008302	0.000100	0.000001	0.000000	0.000000	20
21	0.811215	0.009768	0.000143	0.000003	0.000000	21
22		0.760727	0.011151	0.000196	0.000004	22
23			0.704665	0.012360	0.000256	23
24				0.643940	0.013329	24
25					0.580206	25
$X = 96$						
20	0.007969	0.000092	0.000001	0.000000	0.000000	20
21	0.819516	0.009420	0.000132	0.000002	0.000000	21
22		0.770495	0.010803	0.000182	0.000004	22
23			0.715816	0.012033	0.000240	23
24				0.656300	0.013041	24
25					0.593535	25
$X = 97$						
20	0.007647	0.000084	0.000001	0.000000	0.000000	20
21	0.827486	0.009079	0.000122	0.000002	0.000000	21
22		0.779915	0.010460	0.000169	0.000003	22
23			0.726619	0.011706	0.000224	23
24				0.668333	0.012750	24
25					0.606576	25
$X = 98$						
20	0.007336	0.000077	0.000001	0.000000	0.000000	20
21	0.835133	0.008746	0.000112	0.000002	0.000000	21
22		0.788994	0.010122	0.000156	0.000003	22
23			0.737079	0.011380	0.000209	23
24				0.680039	0.012455	24
25					0.619326	25
$X = 99$						
20	0.007034	0.000070	0.000001	0.000000	0.000000	20
21	0.842469	0.008422	0.000104	0.000002	0.000000	21
22		0.797740	0.009789	0.000145	0.000003	22
23			0.747201	0.011055	0.000195	23
24				0.691418	0.012157	24
25					0.631780	25

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 21$	$R = 22$	$R = 23$	$R = 24$	$R = 25$	K
$X = 100$						
20	0.006742	0.000064	0.000001	0.000000	0.000000	20
21	0.849503	0.008107	0.000095	0.000001	0.000000	21
22		0.806163	0.009463	0.000134	0.000002	22
23			0.756990	0.010734	0.000182	23
24				0.702474	0.011858	24
25					0.643937	25

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
$X = 1$						
1	0.038462	0.037037	0.035714	0.034483	0.033333	1
$X = 2$						
1	0.001479	0.001372	0.001276	0.001189	0.001111	1
2	0.002959	0.002743	0.002551	0.002378	0.002222	2
$X = 3$						
1	0.000057	0.000051	0.000046	0.000041	0.000037	1
2	0.000341	0.000305	0.000273	0.000246	0.000222	2
3	0.000341	0.000305	0.000273	0.000246	0.000222	3
$X = 4$						
3	0.000079	0.000068	0.000059	0.000051	0.000044	3
4	0.000053	0.000045	0.000039	0.000034	0.000030	4
$X = 5$						
$X = 6$						
$X = 7$						
$X = 8$						
$X = 9$						
$X = 10$						
$X = 11$						
$X = 12$						
$X = 13$						
$X = 14$						
$X = 15$						
$X = 16$						
$X = 17$						
$X = 18$						
$X = 19$						
$X = 20$						
$X = 21$						
$X = 22$						
$X = 23$						
$X = 24$						

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
$X = 25$						
$X = 26$						
$X = 27$						
$X = 28$						
$X = 29$						
$X = 30$						
$X = 31$						
$X = 32$						
$X = 33$						
$X = 34$						
$X = 35$						
$X = 36$						
$X = 37$						
$X = 38$						
$X = 39$						
$X = 40$						
$X = 41$						
27		0.000141	0.000032	0.000008	0.000002	27
28			0.000045	0.000011	0.000003	28
29				0.000013	0.000003	29
30					0.000003	30
$X = 42$						
26	0.000616	0.000126	0.000027	0.000006	0.000002	26
27		0.000226	0.000049	0.000011	0.000003	27
28			0.000076	0.000018	0.000004	28
29				0.000024	0.000006	29
30					0.000007	30
$X = 43$						
26	0.000914	0.000180	0.000038	0.000008	0.000002	26
27		0.000352	0.000074	0.000016	0.000004	27
28			0.000125	0.000028	0.000006	28
29				0.000041	0.000010	29
30					0.000012	30
$X = 44$						
25	0.000543	0.000103	0.000021	0.000004	0.000001	25
26	0.001321	0.000251	0.000051	0.000011	0.000002	26
27		0.000533	0.000107	0.000023	0.000005	27
28			0.000199	0.000043	0.000010	28
29				0.000069	0.000015	29
30					0.000022	30
$X = 45$						
25	0.000708	0.000130	0.000025	0.000005	0.000001	25
26	0.001863	0.000341	0.000066	0.000014	0.000003	26
27		0.000784	0.000152	0.000031	0.000007	27

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
28			0.000307	0.000063	0.000014	28
29				0.000112	0.000024	29
30					0.000037	30
$X = 46$						
25	0.000907	0.000160	0.000030	0.000006	0.000001	25
26	0.002571	0.000453	0.000085	0.000017	0.000004	26
27		0.001125	0.000211	0.000042	0.000009	27
28			0.000459	0.000091	0.000019	28
29				0.000175	0.000037	29
30					0.000061	30
$X = 47$						
25	0.001140	0.000193	0.000035	0.000007	0.000001	25
26	0.003478	0.000590	0.000107	0.000021	0.000004	26
27		0.001578	0.000285	0.000055	0.000011	27
28			0.000670	0.000129	0.000026	28
29				0.000266	0.000054	29
30					0.000098	30
$X = 48$						
25	0.001410	0.000230	0.000040	0.000007	0.000001	25
26	0.004618	0.000755	0.000132	0.000024	0.000005	26
27		0.002168	0.000378	0.000070	0.000014	27
28			0.000955	0.000177	0.000035	28
29				0.000395	0.000077	29
30					0.000152	30
$X = 49$						
25	0.001717	0.000270	0.000045	0.000008	0.000002	25
26	0.006027	0.000948	0.000159	0.000029	0.000005	26
27		0.002922	0.000491	0.000088	0.000017	27
28			0.001333	0.000239	0.000045	28
29				0.000573	0.000108	29
30					0.000229	30
$X = 50$						
24	0.000480	0.000073	0.000012	0.000002	0.000000	24
25	0.002062	0.000312	0.000051	0.000009	0.000002	25
26	0.007744	0.001173	0.000190	0.000033	0.000006	26
27		0.003871	0.000628	0.000109	0.000020	27
28			0.001824	0.000316	0.000058	28
29				0.000812	0.000148	29
30					0.000338	30
$X = 51$						
24	0.000533	0.000078	0.000012	0.000002	0.000000	24
25	0.002445	0.000357	0.000056	0.000009	0.000002	25
26	0.009807	0.001431	0.000224	0.000037	0.000007	26
27		0.005044	0.000789	0.000132	0.000023	27
28			0.002452	0.000410	0.000073	28
29				0.001128	0.000199	29
30					0.000486	30
$X = 52$						
24	0.000586	0.000082	0.000012	0.000002	0.000000	24

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
25	0.002863	0.000402	0.000061	0.000010	0.000002	25
26	0.012251	0.001721	0.000260	0.000042	0.000007	26
27		0.006475	0.000976	0.000158	0.000027	27
28			0.003240	0.000523	0.000089	28
29				0.001538	0.000263	29
30					0.000685	30
$X = 53$						
24	0.000637	0.000086	0.000013	0.000002	0.000000	24
25	0.003316	0.000449	0.000065	0.000010	0.000002	25
26	0.015114	0.002045	0.000297	0.000046	0.000008	26
27		0.008197	0.001191	0.000186	0.000031	27
28			0.004216	0.000658	0.000109	28
29				0.002062	0.000340	29
30					0.000948	30
$X = 54$						
24	0.000686	0.000089	0.000013	0.000002	0.000000	24
25	0.003801	0.000495	0.000069	0.000010	0.000002	25
26	0.018430	0.002401	0.000337	0.000051	0.000008	26
27		0.010242	0.001436	0.000216	0.000035	27
28			0.005408	0.000814	0.000130	28
29				0.002719	0.000434	29
30					0.001289	30
$X = 55$						
24	0.000733	0.000092	0.000012	0.000002	0.000000	24
25	0.004314	0.000541	0.000073	0.000011	0.000002	25
26	0.022231	0.002789	0.000377	0.000055	0.000008	26
27		0.012643	0.001709	0.000249	0.000038	27
28			0.006843	0.000995	0.000154	28
29				0.003534	0.000545	29
30					0.001723	30
$X = 56$						
24	0.000776	0.000094	0.000012	0.000002	0.000000	24
25	0.004853	0.000586	0.000076	0.000011	0.000002	25
26	0.026545	0.003207	0.000418	0.000059	0.000009	26
27		0.015432	0.002011	0.000282	0.000042	27
28			0.008552	0.001201	0.000179	28
29				0.004529	0.000676	29
30					0.002268	30
$X = 57$						
24	0.000816	0.000095	0.000012	0.000002	0.000000	24
25	0.005412	0.000630	0.000079	0.000011	0.000002	25
26	0.031398	0.003653	0.000459	0.000062	0.000009	26
27		0.018640	0.002343	0.000318	0.000046	27
28			0.010564	0.001432	0.000207	28
29				0.005730	0.000826	29
30					0.002944	30
$X = 58$						
24	0.000852	0.000095	0.000012	0.000002	0.000000	24
25	0.005989	0.000671	0.000081	0.000011	0.000001	25
26	0.036810	0.004124	0.000500	0.000065	0.000009	26

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
27		0.022293	0.002702	0.000354	0.000049	27
28			0.012907	0.001689	0.000236	28
29				0.007161	0.000998	29
30					0.003770	30
$X = 59$						
24	0.000884	0.000095	0.000011	0.000001	0.000000	24
25	0.006577	0.000710	0.000083	0.000010	0.000001	25
26	0.042798	0.004618	0.000540	0.000068	0.000009	26
27		0.026417	0.003087	0.000390	0.000053	27
28			0.015608	0.001972	0.000266	28
29				0.008851	0.001193	29
30					0.004769	30
$X = 60$						
24	0.000911	0.000095	0.000011	0.000001	0.000000	24
25	0.007174	0.000745	0.000084	0.000010	0.000001	25
26	0.049376	0.005130	0.000578	0.000071	0.000009	26
27		0.031035	0.003497	0.000427	0.000056	27
28			0.018696	0.002281	0.000297	28
29				0.010823	0.001410	29
30					0.005961	30
$X = 61$						
24	0.000933	0.000093	0.000010	0.000001	0.000000	24
25	0.007774	0.000778	0.000085	0.000010	0.000001	25
26	0.056549	0.005658	0.000615	0.000072	0.000009	26
27		0.036165	0.003930	0.000463	0.000058	27
28			0.022193	0.002614	0.000329	28
29				0.013104	0.001650	29
30					0.007372	30
$X = 62$						
24	0.000951	0.000092	0.000010	0.000001	0.000000	24
25	0.008372	0.000807	0.000085	0.000010	0.000001	25
26	0.064323	0.006197	0.000649	0.000074	0.000009	26
27		0.041823	0.004382	0.000499	0.000061	27
28			0.026123	0.002971	0.000362	28
29				0.015719	0.001914	29
30					0.009022	30
$X = 63$						
24	0.000964	0.000089	0.000009	0.000001	0.000000	24
25	0.008965	0.000832	0.000084	0.000009	0.000001	25
26	0.072695	0.006745	0.000681	0.000075	0.000009	26
27		0.048020	0.004852	0.000533	0.000063	27
28			0.030505	0.003350	0.000394	28
29				0.018690	0.002200	29
30					0.010936	30
$X = 64$						
24	0.000973	0.000087	0.000008	0.000001	0.000000	24
25	0.009547	0.000853	0.000083	0.000009	0.000001	25
26	0.081660	0.007296	0.000711	0.000075	0.000009	26
27		0.054765	0.005336	0.000566	0.000064	27
28			0.035357	0.003749	0.000426	28

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
29				0.022040	0.002507	29
30					0.013135	30
$X = 65$						
24	0.000978	0.000084	0.000008	0.000001	0.000000	24
25	0.010116	0.000870	0.000082	0.000008	0.000001	25
26	0.091207	0.007847	0.000737	0.000076	0.000008	26
27		0.062061	0.005831	0.000597	0.000066	27
28			0.040693	0.004166	0.000458	28
29				0.025790	0.002836	29
30					0.015642	30
$X = 66$						
24	0.000978	0.000081	0.000007	0.000001	0.000000	24
25	0.010666	0.000884	0.000080	0.000008	0.000001	25
26	0.101322	0.008394	0.000760	0.000075	0.000008	26
27		0.069908	0.006333	0.000626	0.000066	27
28			0.046523	0.004599	0.000489	28
29				0.029956	0.003184	29
30					0.018478	30
$X = 67$						
24	0.000974	0.000078	0.000007	0.000001	0.000000	24
25	0.011196	0.000893	0.000078	0.000007	0.000001	25
26	0.111989	0.008935	0.000780	0.000075	0.000008	26
27		0.078302	0.006840	0.000653	0.000067	27
28			0.052856	0.005045	0.000518	28
29				0.034555	0.003550	29
30					0.021662	30
$X = 68$						
24	0.000967	0.000074	0.000006	0.000001	0.000000	24
25	0.011702	0.000899	0.000076	0.000007	0.000001	25
26	0.123185	0.009464	0.000797	0.000074	0.000007	26
27		0.087237	0.007348	0.000677	0.000067	27
28			0.059697	0.005502	0.000546	28
29				0.039600	0.003933	29
30					0.025213	30
$X = 69$						
24	0.000956	0.000071	0.000006	0.000001	0.000000	24
25	0.012181	0.000901	0.000073	0.000007	0.000001	25
26	0.134887	0.009979	0.000810	0.000072	0.000007	26
27		0.096701	0.007855	0.000699	0.000067	27
28			0.067045	0.005966	0.000573	28
29				0.045102	0.004330	29
30					0.029146	30
$X = 70$						
24	0.000942	0.000067	0.000005	0.000000	0.000000	24
25	0.012632	0.000900	0.000070	0.000006	0.000001	25
26	0.147069	0.010477	0.000821	0.000071	0.000007	26
27		0.106680	0.008356	0.000718	0.000067	27
28			0.074900	0.006435	0.000597	28
29				0.051068	0.004739	29
30					0.033476	30

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
$X = 71$						
24	0.000925	0.000063	0.000005	0.000000	0.000000	24
25	0.013052	0.000895	0.000068	0.000006	0.000001	25
26	0.159701	0.010956	0.000827	0.000069	0.000006	26
27		0.117157	0.008848	0.000734	0.000066	27
28			0.083255	0.006907	0.000619	28
29				0.057503	0.005158	29
30					0.038215	30
$X = 72$						
24	0.000906	0.000060	0.000004	0.000000	0.000000	24
25	0.013439	0.000888	0.000065	0.000005	0.000000	25
26	0.172753	0.011413	0.000831	0.000067	0.000006	26
27		0.128114	0.009330	0.000748	0.000065	27
28			0.092104	0.007377	0.000639	28
29				0.064410	0.005585	29
30					0.043373	30
$X = 73$						
24	0.000885	0.000056	0.000004	0.000000	0.000000	24
25	0.013794	0.000878	0.000062	0.000005	0.000000	25
26	0.186192	0.011845	0.000832	0.000064	0.000005	26
27		0.139526	0.009798	0.000758	0.000063	27
28			0.101434	0.007845	0.000657	28
29				0.071787	0.006017	29
30					0.048958	30
$X = 74$						
24	0.000862	0.000053	0.000004	0.000000	0.000000	24
25	0.014114	0.000865	0.000059	0.000004	0.000000	25
26	0.199986	0.012251	0.000830	0.000062	0.000005	26
27		0.151371	0.010251	0.000766	0.000062	27
28			0.111232	0.008306	0.000673	28
29				0.079632	0.006451	29
30					0.054974	30
$X = 75$						
24	0.000837	0.000049	0.000003	0.000000	0.000000	24
25	0.014400	0.000850	0.000055	0.000004	0.000000	25
26	0.214100	0.012630	0.000825	0.000060	0.000005	26
27		0.163623	0.010684	0.000771	0.000060	27
28			0.121483	0.008759	0.000686	28
29				0.087939	0.006886	29
30					0.061426	30
$X = 76$						
24	0.000811	0.000046	0.000003	0.000000	0.000000	24
25	0.014651	0.000832	0.000052	0.000004	0.000000	25
26	0.228500	0.012981	0.000817	0.000057	0.000004	26
27		0.176253	0.011098	0.000773	0.000058	27
28			0.132167	0.009202	0.000696	28
29				0.096698	0.007320	29
30					0.068312	30
$X = 77$						
24	0.000784	0.000043	0.000003	0.000000	0.000000	24

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
25	0.014867	0.000814	0.000049	0.000003	0.000000	25
26	0.243151	0.013302	0.000808	0.000054	0.000004	26
27		0.189234	0.011490	0.000773	0.000057	27
28			0.143265	0.009631	0.000704	28
29				0.105900	0.007749	29
30					0.075632	30
$X = 78$						
24	0.000757	0.000040	0.000002	0.000000	0.000000	24
25	0.015050	0.000793	0.000046	0.000003	0.000000	25
26	0.258018	0.013592	0.000796	0.000052	0.000004	26
27		0.202535	0.011858	0.000770	0.000054	27
28			0.154755	0.010045	0.000710	28
29				0.115530	0.008171	29
30					0.083380	30
$X = 79$						
24	0.000728	0.000037	0.000002	0.000000	0.000000	24
25	0.015198	0.000771	0.000044	0.000003	0.000000	25
26	0.273068	0.013853	0.000782	0.000049	0.000003	26
27		0.216128	0.012202	0.000765	0.000052	27
28			0.166613	0.010442	0.000714	28
29				0.125575	0.008586	29
30					0.091552	30
$X = 80$						
24	0.000700	0.000034	0.000002	0.000000	0.000000	24
25	0.015314	0.000748	0.000041	0.000003	0.000000	25
26	0.288266	0.014082	0.000767	0.000047	0.000003	26
27		0.229980	0.012520	0.000759	0.000050	27
28			0.178815	0.010821	0.000715	28
29				0.136018	0.008989	29
30					0.100137	30
$X = 81$						
24	0.000671	0.000032	0.000002	0.000000	0.000000	24
25	0.015398	0.000725	0.000038	0.000002	0.000000	25
26	0.303581	0.014281	0.000750	0.000044	0.000003	26
27		0.244062	0.012813	0.000750	0.000048	27
28			0.191336	0.011181	0.000714	28
29				0.146839	0.009381	29
30					0.109127	30
$X = 82$						
24	0.000643	0.000029	0.000001	0.000000	0.000000	24
25	0.015451	0.000700	0.000036	0.000002	0.000000	25
26	0.318979	0.014450	0.000732	0.000042	0.000003	26
27		0.258344	0.013078	0.000739	0.000046	27
28			0.204149	0.011519	0.000711	28
29				0.158020	0.009759	29
30					0.118507	30
$X = 83$						
24	0.000614	0.000027	0.000001	0.000000	0.000000	24
25	0.015475	0.000676	0.000033	0.000002	0.000000	25
26	0.334430	0.014590	0.000712	0.000039	0.000002	26

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
27		0.272794	0.013317	0.000727	0.000044	27
28			0.217227	0.011835	0.000707	28
29				0.169538	0.010121	29
30					0.128266	30
$X = 84$						
24	0.000586	0.000025	0.000001	0.000000	0.000000	24
25	0.015470	0.000650	0.000031	0.000002	0.000000	25
26	0.349905	0.014700	0.000692	0.000037	0.000002	26
27		0.287384	0.013528	0.000713	0.000041	27
28			0.230543	0.012129	0.000700	28
29				0.181373	0.010467	29
30					0.138387	30
$X = 85$						
24	0.000558	0.000023	0.000001	0.000000	0.000000	24
25	0.015439	0.000625	0.000028	0.000002	0.000000	25
26	0.365376	0.014782	0.000671	0.000034	0.000002	26
27		0.302083	0.013712	0.000698	0.000039	27
28			0.244071	0.012399	0.000692	28
29				0.193502	0.010795	29
30					0.148854	30
$X = 86$						
24	0.000531	0.000021	0.000001	0.000000	0.000000	24
25	0.015382	0.000600	0.000026	0.000001	0.000000	25
26	0.380814	0.014836	0.000650	0.000032	0.000002	26
27		0.316865	0.013870	0.000682	0.000037	27
28			0.257784	0.012645	0.000682	28
29				0.205900	0.011104	29
30					0.159648	30
$X = 87$						
24	0.000505	0.000019	0.000001	0.000000	0.000000	24
25	0.015301	0.000575	0.000024	0.000001	0.000000	25
26	0.396196	0.014864	0.000628	0.000030	0.000002	26
27		0.331701	0.014001	0.000665	0.000035	27
28			0.271653	0.012867	0.000671	28
29				0.218545	0.011393	29
30					0.170752	30
$X = 88$						
24	0.000479	0.000017	0.000001	0.000000	0.000000	24
25	0.015198	0.000550	0.000023	0.000001	0.000000	25
26	0.411497	0.014867	0.000606	0.000028	0.000001	26
27		0.346565	0.014106	0.000647	0.000033	27
28			0.285654	0.013066	0.000659	28
29				0.231413	0.011663	29
30					0.182145	30
$X = 89$						
24	0.000454	0.000016	0.000001	0.000000	0.000000	24
25	0.015074	0.000525	0.000021	0.000001	0.000000	25
26	0.426695	0.014846	0.000583	0.000026	0.000001	26
27		0.361432	0.014187	0.000629	0.000031	27
28			0.299761	0.013240	0.000646	28

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
29				0.244478	0.011911	29
30					0.193807	30
$X = 90$						
25	0.014930	0.000501	0.000019	0.000001	0.000000	25
26	0.441769	0.014802	0.000561	0.000024	0.000001	26
27		0.376278	0.014242	0.000610	0.000029	27
28			0.313947	0.013390	0.000632	28
29				0.257718	0.012139	29
30					0.205719	30
$X = 91$						
25	0.014769	0.000477	0.000018	0.000001	0.000000	25
26	0.456699	0.014736	0.000539	0.000023	0.000001	26
27		0.391079	0.014275	0.000590	0.000027	27
28			0.328190	0.013517	0.000617	28
29				0.271108	0.012345	29
30					0.217857	30
$X = 92$						
25	0.014592	0.000454	0.000016	0.000001	0.000000	25
26	0.471468	0.014650	0.000516	0.000021	0.000001	26
27		0.405815	0.014284	0.000571	0.000026	27
28			0.342464	0.013621	0.000601	28
29				0.284625	0.012530	29
30					0.230202	30
$X = 93$						
25	0.014399	0.000432	0.000015	0.000001	0.000000	25
26	0.486060	0.014545	0.000495	0.000019	0.000001	26
27		0.420465	0.014272	0.000551	0.000024	27
28			0.356748	0.013702	0.000585	28
29				0.298246	0.012693	29
30					0.242732	30
$X = 94$						
25	0.014194	0.000410	0.000014	0.000001	0.000000	25
26	0.500459	0.014422	0.000473	0.000018	0.000001	26
27		0.435010	0.014239	0.000531	0.000022	27
28			0.371020	0.013761	0.000568	28
29				0.311947	0.012836	29
30					0.255426	30
$X = 95$						
25	0.013976	0.000389	0.000012	0.000001	0.000000	25
26	0.514653	0.014283	0.000452	0.000017	0.000001	26
27		0.449432	0.014187	0.000511	0.000021	27
28			0.385259	0.013799	0.000551	28
29				0.325709	0.012958	29
30					0.268262	30
$X = 96$						
25	0.013747	0.000369	0.000011	0.000000	0.000000	25
26	0.528629	0.014128	0.000431	0.000015	0.000001	26
27		0.463715	0.014116	0.000491	0.000020	27
28			0.399446	0.013817	0.000534	28

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

K	$R = 26$	$R = 27$	$R = 28$	$R = 29$	$R = 30$	K
29				0.339508	0.013059	29
30					0.281219	30
$X = 97$						
25	0.013509	0.000349	0.000010	0.000000	0.000000	25
26	0.542376	0.013960	0.000411	0.000014	0.000001	26
27		0.477844	0.014027	0.000472	0.000018	27
28			0.413562	0.013815	0.000517	28
29				0.353325	0.013140	29
30					0.294278	30
$X = 98$						
25	0.013263	0.000330	0.000010	0.000000	0.000000	25
26	0.555886	0.013779	0.000391	0.000013	0.000001	26
27		0.491804	0.013923	0.000453	0.000017	27
28			0.427589	0.013794	0.000499	28
29				0.367140	0.013202	29
30					0.307418	30
$X = 99$						
25	0.013010	0.000312	0.000009	0.000000	0.000000	25
26	0.569149	0.013587	0.000372	0.000012	0.000001	26
27		0.505583	0.013803	0.000434	0.000016	27
28			0.441512	0.013756	0.000482	28
29				0.380934	0.013244	29
30					0.320620	30
$X = 100$						
25	0.012751	0.000295	0.000008	0.000000	0.000000	25
26	0.582159	0.013384	0.000354	0.000011	0.000001	26
27		0.519170	0.013669	0.000415	0.000015	27
28			0.455315	0.013700	0.000465	28
29				0.394690	0.013269	29
30					0.333864	30

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

GLOSSARY OF STATISTICAL TERMS[1]

Average

A measure of location or position of a frequency distribution. There are three types of average in general use: the mean (arithmetic, geometric and harmonic), the median and the mode.

Binomial Distribution

If the probability of an event to appear in any one trial is p, and not to appear $q = 1 - p$, then provided that the chance p (q) remains constant throughout the trials, the expected frequencies of 0, 1, 2 ... appearances in N sets of n trials are the 1st, 2nd, etc. terms in the expansion of the binomial

$$N(q + p)^n$$

For r appearances of the event in n independent trials, the general term of the binomial is

$$\binom{r}{n} q^{n-r} p^r$$

The distribution is also known as the Bernoulli distribution, after James Bernoulli, *Ars conjectandi*, 1713. The mean of the binomial is pn and its standard deviation is $\sqrt{npq}$.

Binomial Index of Dispersion

A statistic for testing the homogeneity of a set of samples with respect to some common attribute.

If there are k samples of sizes $n_1, n_2 \ldots n_k$ respectively, and p is the mean proportion for all sets together, i.e.

$$p = \left(\sum_{i=1}^{k} n_i p_i\right) \Big/ \sum_{i=1}^{k} n_i$$

[1] For some of the definitions I am indebted to the *Dictionary of Statistical Terms* by M. G. Kendall & W. R. Buckland (London, 1957).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the index of dispersion is

$$\sum_{i=1}^{k} n_i (p_i - p)^2 / p(1-p)$$

The significance of the index, as denoting departure from homogeneity, is tested by the χ^2 distribution with $k - 1$ degrees of freedom.

The index is a particular case of the Lexis ratio (q.v.)

Correlation

In its most general sense correlation denotes the interdependence of quantitative or qualitative data. This concept may be extended to more than two variates.

In a narrower sense it denotes the relationship between measurable variates or ranks.

Correlation, coefficient of

The correlation coefficient is a measure of the interdependence between two or more variates. It is usually a pure number which may vary between -1 and 1 with the intermediate value of zero indicating absence of correlation. The limiting values indicate complete correlation, positive or negative, and thus functional relationships.

The most commonly used measure of dependence is Pearson's Product-Moment Coefficient of Correlation, whose numerator is the first product-moment or Covariance (the sum of joint occurrence of the two variates) and whose denominator is the square root of the product of the sums of squared deviations of the two variates:

$$S = \frac{\Sigma(xy)}{\sqrt{S(x^2)\ S(y^2)}}$$

where x, y are the values of the variables measured from their respective means.

The Product-Moment Correlation Coefficient includes the so-called Bernoullian (tetrachoric) correlation coefficient in which each variable can take only two values (Sect. 6.2), and Spearman's coefficient of rank correlation if the ranks are used as values of the variable (Vol. I, Sect. 7.3).

Distribution

In the descriptive sense, the splitting up of a mass of elements among a number of classes; in the mathematical sense, the function governing such division of masses.

Estimator, unbiased

An estimator, say, m, which for all sample sizes has its expected value equal to the parameter under estimate, say, μ; $E(m) = \mu$. The statistic $\frac{1}{n-1}\Sigma(x_i - \bar{x})^2$ s an unbiassed estimate of the population variance since $E\left[\frac{1}{n-1}\Sigma(x_i - \bar{x})^2\right] = \sigma$.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Estimator, consistent

An estimator which converges in probability as the sample size increases to the parameter of which it is an estimator.

Expectation

The expected value of a statistic is its mean value in repeated sampling.

Exponential Curves

A series of observations ordered in time, which has a constant or approximately constant rate of increase may be represented over a long period by the curve

$$y = ae^{bt}$$

where a and b are constants and t is time. This is called the exponential curve. The fitting of an exponential trend of this form by the method of least squares is facilitated by transforming into the logarithmic form

$$\log_e y = \log_e a + bt$$

Frequency

The number of occurrences of a given type of event, or the number of members of a population falling into a specified class.

Where the frequency is expressed as a proportion of the total number of occurrences, it is called the *relative* or *proportional* frequency. These ratios are often simply called 'frequencies'.

Frequency Curve

A graphical representation of a continuous frequency distribution, the variate being the abscissa and the frequency the ordinate. It may be fitted as the limiting form of the frequency polygon as the number of observations tends to become infinitely large, and the class intervals infinitely small.

F-Test

An alternative name for the Variance-Ratio-Test. The test function is based upon the mathematical fact that the variance ratio of samples from a normal population depends only upon the degrees of freedom in the two samples, and is independent of the particular value of the standard deviation of individuals in the normal population.

Factor Analysis

A branch of multi-variate analysis in which the observed variates x_i $(i = 1, 2 \ldots p)$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

are supposed to be expressible in terms of number $m<p$ factors f_j together with residual elements. One such model is expressed by

$$x_i = \sum_{j=1}^{m} a_{ij} f_j + b_i s_i + c_i \varepsilon_i$$

where s_i is a factor specific to the i'th variate.

ε_i is an error term and the a's, b's and c's are constants of the model which is the object of the analysis to estimate. The coefficients a_{ij} are known as *loadings* of the factors. That part of the variance of x_i which is attributable to the f's is called the *communality*; that attributable to s_i is called the *specificity*, and that attributable to ε_i is called the *unreliability*.

The expressions "component analysis" and "factor analysis" occur also in variance analysis with a different meaning, namely in relation to the allocation of variance to different causal factors or components of variation.

Fluctuation

The variation of a statistic from sample to sample is referred to as a sampling fluctuation

Lexis' Ratio

Consider a number of samples of n individuals in which the proportion of elements having a certain attribute are $p_1, p_2 \ldots p_k$; the nature of the universe from which they were drawn is to be ascertained—whether it is homogeneous and the samples may be regarded as obtained by simple (random) sampling (q.v.), whether it varies with time or place so that the samples are not simple, and so on. If p is the mean of the observed $p_1 \ldots p_k$, we write for the probable error of p in a sample of size n, according to the Bernoullian law

$$r = \frac{pq}{\sqrt{n}}$$

and if calculated without the assumption of that law

$$R = \frac{S(p_k - p)^2}{\sqrt{k-1}}$$

where the summation extends over all values $p_1, p_2 \ldots p_k$, and define the Lexis' ratio as

$$Q = R/r$$

If we work with the absolute frequencies, the formulae become

$$r = \sqrt{npq}$$

$$R = \frac{S(n_k - \bar{n})^2}{\sqrt{k-1}}$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

If the samples are random samples from a homogeneous universe (the case of simple sampling), R is for large samples approximately equal to r, and $Q = 1$. This case Lexis calls "normal dispersion".

On the other hand, if the universe is not constant while the samples are drawn, or if they come from different parts of a patchy universe, R is no longer an estimate of the probable error of the constant p, but may be split into two parts, one due to the sampling fluctuation of p round the mean value (the aleatoric part), the other due to the variation of the true values round that mean (the physical part). R will therefore be greater than r and $Q > 1$. This case Lexis calls "super-normal dispersion", and the universe from which such samples are drawn is known as a Lexis' universe.

A mass event with a constant probability leads to a Q indicative of normal dispersion, and a mass event with a variable probability to a Q indicative of super-normal dispersion, and since another possibility of connecting independent chance events is unthinkable, it follows that a Q which is significantly less than unity does no longer refer to pure chance events, but must be indicative of a linkage between the individual events. This case Lexis calls "sub-normal dispersion", and the universe which gives rise to it is characterised by "constraints".

The quantity Q is analogous to the quantity χ^2.

Level of Significance

Many statistical texts of hypotheses depend upon the use of probability distributions of a statistic t. The values of the statistic t are used as a criterion of the acceptability of the hypothesis, if their probability P is small in the sense of falling outside the range defined by the specified values of t, t_o and t_1 ($t_1 > t_o$) the hypothesis is rejected. The probabilities $P(t > t_1)$ and $P(t \geqslant t_o)$ are called levels of significance, and are usually expressed as percentages. The choice of level is arbitrary, but popular values are 5, 1 and 0.1 per cent. Thus, for example, if the probability of all values $> t_1$ or $< t_o$ is 0.01, any observed value of t beyond t_1 or t_o would make us reject the hypothesis with a probability 1 in 100 of being correct, because there is still a probability of 1 in 100 of $t > t_1$ and $t < t_o$, even if the hypothesis is correct.

Lognormal Distribution

If the logarithms of a set of variate values are distributed according to the normal distribution, the variate is said to have a logarithmic-normal distribution, or to be distributed 'lognormally'.

Mean, Arithmetic

The arithmetic mean (M or x) of n values $x_1, x_2 \ldots x_n$ is given by

$$M = \frac{1}{n} S(x)$$

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

where $S(x)$ denotes "the sum of all quantities like x". If the values $x_1, x_2 \ldots x_n$ occur with frequencies $f_1, f_2 \ldots f_n$ respectively, the formula becomes

$$M = \frac{1}{n} \overset{n}{\underset{1}{S}}(f_t x_t)$$

where x_t denotes any value of x, and f_t the corresponding frequency; n is the sum of all frequencies or $S(f_t)$. It may be briefly denoted by S_o and $S(f_t x_t)$ by S_1. In this notation the arithmetic mean is

$$M = S_1/S_o$$

Mean, Geometric

The geometric mean is given by

$$G = (x_t \ldots x_n)\ 1/n$$
$$\log G = 1/n\ S(\log x)$$

from which it follows that the logarithm of the geometric mean G of a series of values is the arithmetic mean of their logarithms. An approximate formula if the deviations are small compared with M is $G = M\left(1 - \frac{\sigma^2}{2M^2}\right)$.

Mean, Harmonic

The harmonic mean of a series of quantities is the reciprocal of the arithmetic mean of the reciprocals; if H be the harmonic mean

$$1/H = \frac{1}{n} S\left(\frac{1}{x}\right)$$

An approximate formula if the deviations are small compared with M is

$$H = M\left(1 - \frac{\sigma^2}{M^2}\right)$$

Mean Deviation

The mean deviation is the arithmetic mean of the "absolute" deviations d from the average

$$m.d. = \frac{1}{n} S(d)$$

It is a minimum about the median.

Median

The median is the middlemost or central value of the variable, that is, the value which is equally likely to exceed other values of the variable than be exceeded by them when

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji.

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

the values are ranged in order of magnitude. If the number of values is even, the median is conventionally taken to be the arithmetic mean of the two central values.

The Mode

The mode is the value of the variate corresponding to the maximum or peak of the distribution curve.

Moment

In general, the moment is the mean value of a power of the variate; in particular, the rth moment of a variate x with distribution $dF(x)$ is given by

$$\mu'_r = \int_{-\infty}^{\infty} x^r dF(x)$$

and the moment about a particular fixed value a, e.g. the arithmetic mean, by

$$\int_{-\infty}^{\infty} (x-a)^r \, dF(x)$$

Moments, Methods of

A method of curve-fitting which proceeds by identifying the lower moments of the observed data with those of the particular curve form to be fitted.

Normal Distribution

The continuous frequency distribution represented by the equation:

$$dF = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} dx \quad \Big| \quad -\infty \leqslant x \leqslant \infty$$

where m is the mean, and σ the standard deviation.

The distribution is often referred to as the Gaussian, the Laplacean, or the Second Law of Laplace. It was apparently first discovered by De Moivre (1756) as the limiting form of the binomial distribution.

Normal law, Reproductive Property of

An aggregate deviation, formed by the summation of any number of deviations which obey the Normal Law, itself obeys the Normal Law. This is the reproductive property, or group property, of the Normal Law of Frequency. It is by virtue of it that the means of samples of a specified size withdrawn at random from a normal population are again normally distributed; also that the differences of either individual values or mean values from such samples are normally distributed.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

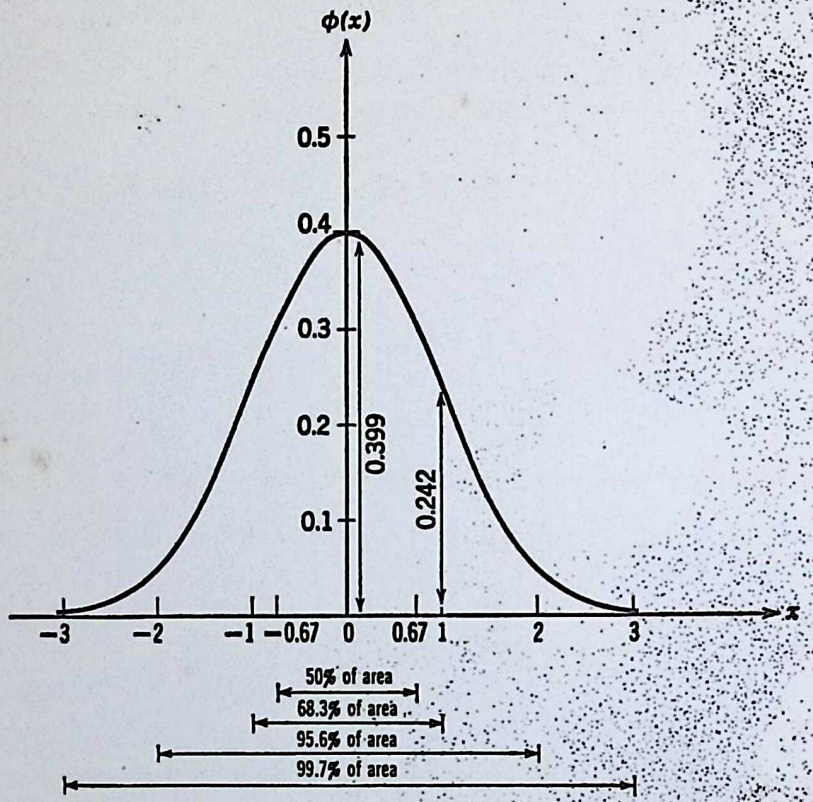


FIG. 16. The normal curve

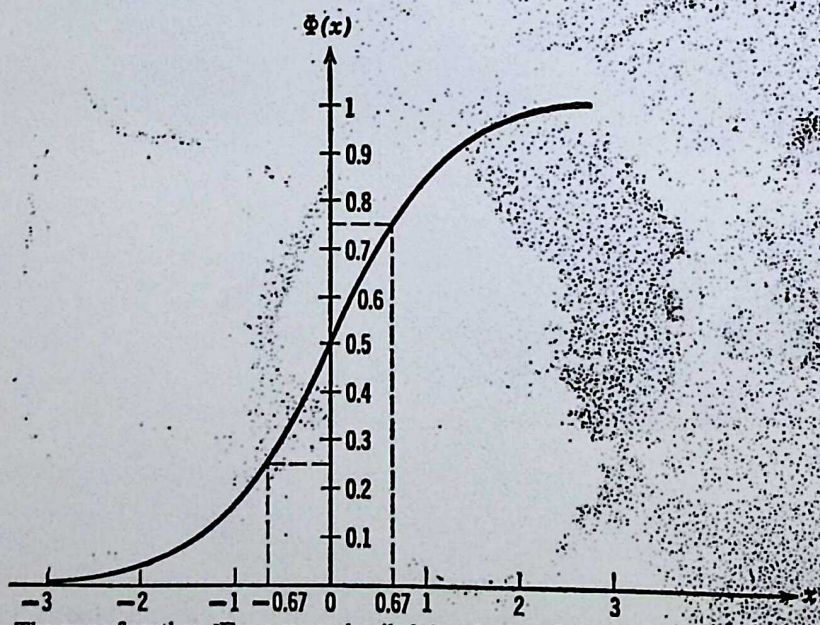


FIG. 17. The error function. [For a more detailed description of the properties of these curves see Vol. I, Sect. 16.4]

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC-0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Probit

The probit corresponding to a given percentage is the normal deviate (increased by 5 to avoid negative values) of the *error function*, that is the cumulative or integral of the normal distribution.

The relation between the Normal Law and its integral (the error function) is shown in Figs. 16, 17.

Sample

A part of a population or universe of units.

Sample Size

The number of sampling units which are included in the sample.

Sampling Distributions

The distribution of a statistic (as different from the sampling unit or individual) in all possible samples which can be chosen according to some specified sampling scheme.

Sampling Error

That part of the difference between a population value characterising a given universe and an estimate thereof obtained from a random sample, which is due to the fact that only a sample of values is observed; as distinct from errors due to bias in selection, errors of observation and recording etc.

Sampling Unit

One of the units into which an aggregate is divided for the purpose of sampling, each unit being regarded as individual and indivisible for the purpose of selection.

Significance

A difference between units or statistics is said to be significant (i.e. indicative of an assignable, "real", cause) if it lies outside acceptable limits of random variation.

What is regarded as "acceptable" depends upon how reliable we wish to make our conclusions.

Standard Deviation

The standard deviation σ is a measure of the scatter or dispersion of the distribution about the mean. If $x_1, x_2 \ldots x_n$ are the deviations of $X_1, X_2 \ldots X_n$ from the Mean M, σ is defined as the root-mean square deviation from the mean by

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

$$\sigma^2 = \frac{1}{n} S(x_t^2)$$

where x_t is any deviation from the arithmetic mean. σ^2 is called the "variance".

For a grouped distribution in which every value of the variable may be represented by a number of occurrences, $f_1, f_2 \ldots f_n$, the formula becomes

$$\sigma^2 = \frac{1}{n} S(f_t x_t^2)$$

Between the sum of squared values of the variable $S(X_t^2)$ and the sum of their deviations from an arbitrary origin A, $S(x_{tA})^2$, there exists the inequality relation

$$S(X_t^2) \geqslant S(x_{tA})^2$$

the difference being nA^2.

If A is the arithmetic mean M, we have, therefore, for the standard deviation

$$n\sigma^2 = S(x_{tM}{}^2) = S(X_t^2) - nM^2$$

and

$$\sigma^2 = S(X_t^2)/n - M^2$$

Denoting $S(X)$ by S_1 and $S(X_t^2)$ by S_2, we have

$$\sigma^2 = S_2/n - (S_1/n)^2$$

which is, as a rule, used for the calculation of σ.

The sums S_1 and S_2 are sometimes called the first and second moments of the distribution about zero as origin.

Statistic

A summary value calculated from a sample of observations; usually, but not necessarily, an estimator of some population parameter.

Statistics

As a plural, numerical data (tabulation) relating to an aggregate of individual observations; as a singular, the science of collecting, analysing and interpreting such data.

Variable

A quantity which may take any one of a set of specified values. The same word is applied to denote non-measurable characteristics, which are subdivided into qualitatively different categories.

Variance Analysis

The total variation displayed by a set of observations, as measured by the sums of squares of deviations from the mean, may, under certain conditions, be separated

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

into components associated with specified sources of variation. Such an analysis is called an Analysis of Variance, although in the strict sense it is an analysis of sums of squares.

Variance ratio test (F-test)

A test based on the ratio of two independent statistics, each of which is distributed as the variance in samples from normal populations with the same parent variance.

Usually the statistics themselves are quadratic estimations of the parent variance, that is squared standard deviations.

The test is widely employed in variance analysis to test the homogeneity of a set of means.

Variate

In contradistinction to variable, a quantity which may take different values with different frequency or probability. Variate is, therefore, another name for Random or Chance Variable.

Variation, Coefficient of

The standard deviation of a distribution σ_x divided by the arithmetic mean x; if multiplied by 100, it expresses the standard deviation as a percentage of the mean

$$v_x = \sigma_x/\bar{x}$$

It enables us to compare the variabilities of different distributions, because as a pure number, it is independent of the scale.

The squared coefficient of variation of the mean in samples of n

$$v_m^2 = \frac{\sigma_x^2/n}{\bar{x}^2}$$

has been shown (Herdan) to approximate for great samples Yule's Characteristic K.

Yule's Characteristic

Originally the parameter devised by Yule for characterising style using the criterion of differential frequency of use of words:

$$K = 10^4 \frac{S_2 - N}{N^2}$$

where

$$N = \Sigma(rn_r)$$

$$S_2 = \Sigma(r^2 n_r)$$

$$n_r = \text{(frequency of vocabulary items used } r \text{ times.}$$

and

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Yule's derivation was based upon the hypothesis of the Poisson law for rare events governing the frequency distribution of r.

Derivations of K which do not assume a particular distribution law are those by Herdan as the relative variance of the mean frequency, and by Good as the repeat rate of vocabulary items.[2]

In formula (1) both r and n_r are frequencies of word occurrences and vocabulary respectively, and may by division through the corresponding total be transformed into relative frequencies. For great samples, these relative frequencies are sensibly stable and may be regarded as estimates of the corresponding probabilities, $p_r = r/N$ and $\pi_r = n_r/n$ $(n = n_1 + n_2 + \ldots n_s)$ respectively.

Using the first transformation, we have

$$K^* = S_2/N^2 = \sum_{r=1}^{s} \frac{n^2}{N^2} n_r = \sum_{r=1}^{s} p_r^2 \, n_r$$

and summing over the individual p_r instead of over their groups n_r, we may write:

$$K^* = \sum_{\mu=1}^{n} p_\mu^{\,2}$$

which shows K to be the repeat rate of particular words.

Using the second transformation we have

$$K - 1/n = S_2/N^2 - 1/n = \frac{n \Sigma r^2 n_r - [\Sigma r n_r]^2}{n [\Sigma r n_r]^2}$$

Dividing numerator and denominator by n^3 we get:

$$K^* = \frac{\{\Sigma r^2 \pi_r - [\Sigma r\pi]^2\}/n}{[\Sigma r\pi]^2} = \frac{\sigma_r^2/n}{M_r^2}$$

which shows K, for great n, to be equal to the coefficient of variation of the mean of the variable r.

[2] The following derivation has been given in different contexts (Sects. 5.3, 14.2), but it was thought advisable to repeat it in the Glossary for quick reference.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji.

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

GLOSSARY OF LINGUISTIC TERMS[1]

Agglutination: In general, the addition to a word or root of suffixes for the purpose of expressing grammatical categories and relationships, syntactical function, etc.

Allophones: The several variants of speech sounds, which constitute a phoneme.

Alphabet: The written characters of a language, in their conventional order, each representing a sound or combination of sounds, exclusive of ideograms or logograms and syllable-signs.

Alveolar: In phonetical terminology, a consonant pronounced with the tongue touching the gum above the upper teeth (e.g. *s*, *z*). Also called post-dental.

Alveolo-palatal: In phonetical terminology, a consonant pronounced with the tongue held between the ridge behind the upper teeth and the front part of the hard palate.

Amalgamating language: A flexional language in which the affixes are intimately fused with the roots of the words and do not possess or retain independent identity.

Anagram: A word, name or phrase formed by transposition of the letters of another.

Analogical extension: The modification of the form or meaning of a word on the basis of analogy with, or imitation of, existing or widely used patterns.

Analogical form: A form or construction made on the pattern of several others, according to a determined rule or law.

Analytic language: A language in which auxiliary words are the chief or sole means of expressing grammatical relationships of words, to the total or partial elimination of inflection.

Aphasia: A broad term, subsuming various losses of speech functions due to cerebral lesions or injuries to other nerve centres. Inability to speak with normal symbolic control.

Article: The designation of auxiliary words, used in many languages as a particle inserted before or prefixed to the noun (in certain languages added to the noun as a suffix) to define, limit or modify its use. Articles are classified mainly as definite, indefinite, and partitive; grammars of certain languages list other types of articles too.

[1] For some of the definitions I am indebted to Mario A. Pei & F. Gaynor, *A Dictionary of Linguistics* (1954), and to J. Whatmough, *Language* (1955).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Articulation: The formation of sounds by the vocal organs; the totality of the movements of the organs of speech required for pronouncing distinct sounds.

Aspect: With reference to verbs, a category indicating whether the action or state denoted by the verb is viewed as completed or in progress, as instantaneous or enduring, as momentary or habitual, etc. Aspects of a verb are variously formed in the various languages, by prefixes, suffixes, infixes, by phonetical changes of the root, by the use of auxiliaries, etc.

Aspirate: In phonetical terminology, (1) the sound /*h*/ and its cognates; (2) in general, a sound the pronunciation of which involves an intensity in the expulsion of air from the oral cavity.

Atomistic approach: That approach to language which considers every phoneme as an individual, isolated unit, in contrast to the structural approach.

Attribute: A word or group of words which denotes a quality or characteristic of a substantive.

Attributive adjective: A descriptive adjective which modifies a noun or noun-equivalent directly (e.g. the *white* house), in contradistinction to predicate adjectives.

Auxiliary: In general, a word which has no complete meaning in itself and is used only in combination with or in reference to another word which has a meaning of its own (e.g. prepositions, conjunctions, etc.)

Auxiliary verb: A verb used in combination with another verb, to express the mood, tense, or aspect of the action denoted by that main verb (e.g. the English, *may*, *shall*, etc.)

Back vowel: A vowel, the point of articulation of which is in the rear part of the oral cavity, i.e. pronounced with the back part of the tongue arched toward the palate.

Basic English: A system, devised by C. K. Ogden, consisting of 850 selected English words and simple rules for using this vocabulary to express ideas. The purpose of this system is two-fold: to serve as a simple and easy international auxiliary language, and as an instrument of semantic training. The vocabulary of 850 words consists of 400 general nouns, 200 names of picturable objects, 150 adjectives, and 100 "operators" (verbal forms, prepositions, particles, etc.); a supplementary general science vocabulary of 100 words and special word-lists for specialised scientific fields are also included in the system, to provide a basic scientific vocabulary in addition to the general vocabulary of 850 words.

Bilabiodental: A sound formed with the upper lip and upper teeth touching the lower lip.

Bilateral opposition: Two phonemes with the same archiphoneme, opposed to each other on the basis of a single relevant feature (the mark of opposition).

Bilingual: (1) With reference to a text, written in two languages; (2) with reference to persons, speaking two languages customarily and with equal ease.

Binary: Consisting of two elements; showing two aspects or characteristics. Said of a system of enumeration which uses only two digits, viz. 1 and 0. Also called dyadic. Contrast denary (or decimal), which uses ten digits, 0, 1, 2, 3 . . . 9.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Borrowed words: A word taken over from another language; also, a word of foreign origin, even though modified in form.

Bound form: Short for *bound linguistic form*: a morpheme which has a distinct meaning only when attached (prefixed or suffixed) to a word (e.g. a declensional ending).

Cadence: The modulation of the voice in speaking, which produces the "flow of the language"; especially, the rise and fall produced by the alternation of louder and softer syllables in accentual tongues. The term is applied in particular to the fall of the voice at the end of a sentence or at a pause in general.

Cardinal vowel: One of a series of vowel sounds, not taken from any specific languages, having a characteristic tongue-position and well-defined acoustic qualities; to be used for comparison with vowels in different languages.

Case: In the flexional languages, a morphological variant of a noun, adjective, pronoun, numeral, or participle, distinguished from other such variants of the same word by a specific declensional ending. As applied to non-flexional languages *case* means in general the grammatical function or syntactical relationship of a word, indicated by a preposition, suffix or a particle, or even by word order alone.

Causal conjunction: A conjunction which introduces a clause expressing a purpose (e.g. in order that) or an effect (e.g., therefore).

Causative case: A declensional case in certain languages (e.g. in Caucasian languages), having the denotation *because of*.

Causative suffix, infix, prefix: A particle added to a verb to express that the subject causes another agent to perform the action.

Causative verb: A verb expressing or indicating that the subject causes an agent to perform the action.

Chinese: A sub-family of the Sino-Tibetan family of languages.

Church Slavonic: The South Slavic language into which Kyrillos and Methodos translated the Gospels in the ninth century A.D.

Class: The totality of linguistically, phonetically, morphologically etc. comparable or equivalent elements.

Clause: In general a subdivision of a sentence containing a subject and a predicate (or a word which implies or replaces a predicate).

Cluster: In phonetics, a group of phonemes, not necessarily constituting a syllable.

Cognate words: Two or more words in different languages which have the same root.

Colloquial style: The language or the style of writing used by educated persons in familiar or informal discourse.

Colloquialism: An informal expression or linguistic form of the conversational type; generally considered out of place in formal discourse or literature, although widely used in daily speech.

Common case: The uninflected form of a word which in uninflected languages is used,

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

with or without prepositions, in all instances where inflectional languages would use different case forms.

Common language: That version of a language which is accepted and employed by various groups or strata of the speakers of the language, as a result of the suppression or levelling of local dialects, among the better educated classes at least.

Comparative grammar: That branch of explanatory grammar which examines and compares the grammatical features and phenomena of two or more related languages, in order to ascertain the nature and degree of their relationship.

Comparative linguistics: That branch of linguistics which discloses and studies similarities between related languages; the study of related languages, with particular emphasis on similarities and differences.

Compound: Consisting of two or more parts, elements or components. (The term *compound* is frequently used as a noun, instead of the term *compound word*.)

Compounding: The process of forming new words (compound words) by combining existing words (called the components).

Concord: The agreement of a word with another word or other words in a sentence as to number, case, gender, person, etc.

Conjugation: The scheme of the modifications of the verb, by means of inflection, prefixes, auxiliaries etc. to express various tenses, moods, voices, persons and numbers.

Conjunction: A word used to connect words or sentences, indicating the relationship of these connected elements.

Consonant: (1) A sound produced by an obstruction or blocking or some other restriction of the free passage of the air, exhaled from the lungs, through the oral cavity. (2) A letter representing such a sound.

Consonantal digraph: A combination of two consonant letters representing one single consonant sound.

Consonantal trigraph: A term occasionally applied to any combination of three consonant letters which represents one consonant sound (e.g. German sch.)

Consonantal writing: A system of alphabetic writing in which only consonants are written.

Construction: The manner of grouping and combining elements of speech; specifically syntactical construction.

Content: The intrinsic meaning of a linguistic symbol.

Context: The sounds, words or phrases immediately preceding or following the sound, word, or phrase under consideration; also, the relation to or connection with such preceding or subsequent sounds, words or phrases.

Declension: In inflexional languages, the system of the inflexion of a noun, pronoun, adjective, or numeral, intended to show differences of case, gender and number.

Defective writing: In the Hebrew system of consonantal writing, the method of writing consonants only.

Definite article: A particle inserted before or prefixed (in certain languages added as a

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

suffix) to nouns to indicate that the noun is used to denote a specific member of the class which it designates.

Dental: In phonetical terminology, a consonant pronounced with the tongue touching the tips of the upper teeth.

Dependent clause: A subordinate clause.

Derivation: (1) The formation of a new word from an existing word, root or stem, by the addition of a prefix or suffix, or by other means. (2) The indication of the sources and development of the elements composing a word.

Derivative: In general, a word formed from another word by internal vowel change ("primary derivative") or by the addition of a prefix or suffix ("secondary derivative").

Descriptive linguistics: Synchronic linguistics.

Diachronic linguistics: The study of words, speech, languages and linguistic changes from the point of view of evolution in the course of time.

Diacritic mark; diacritic sign: A sign placed over, below or across a letter, at times merely as an orthographic sign, but in most instances to indicate the correct sound represented by that letter in the given word. (Often referred to as *written accent* or *graphic accent*.)

Diction: The selection and use of words in speech.

Dictionary: A book containing the words of a given language (usually in alphabetic order) with definitions of their meanings; also a book listing the words of a language with their equivalents in a foreign language or foreign languages.

Diagraph: A combination of two letters (either vowel or consonant signs) representing one spoken sound (e.g. the English *th*).

Diphthong: A union of two vowels pronounced in one syllable.

Discourse: In general, any form of oral or written communication or composition.

Dissyllabic: Consisting of two syllables.

Distinctive function: The use of phonemically relevant features to distinguish linguistic signs.

Distribution: The syntagmatic arrangement of elements in the spoken chain.

Dorsal: A sound made with the dorsum.

Dorsum: The back part of the tongue, associated with the soft palate.

Duration: With reference to speech sounds or phonemes, the relative length of time during which the vocal organs remain in the position required for the articulation of a given sound (synonymous with quantity).

Dynamic linguistics: The linguistic discipline which considers languages from the aspect of their development and history.

Form: A recurring identity of utterance. A form is said to be "free" if it may stand alone and still have meaning, "bound" if not, e.g. kind: kindly: *-ly*.

Formant: A bound form which is added to a root or stem, e.g. *-ie-* and *-tat-* in Latin *societat (-em, -is)*.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Form-class: A group into which forms showing recognisable phonetical or grammatical similarities are classified.

Function: In grammatical terminology, the part played in a sentence by a word or ending.

Functional linguistics: Linguistic study in which phonemes and other linguistic elements are viewed and classified according to their function in the given language.

Gemination: Doubling—especially with reference to consonant sounds. In alphabetic writing, gemination is usually indicated by a double consonant or vowel letter.

Gender: A grammatical classification of forms, usually in nouns, pronouns and adjectives, fulfilling the same function and marked by agreement, e.g. Latin bona femina (feminine) "a good woman", bonus poeta (masculine), "a good poet", not a classification by sex.

Generic term: A non-specific, non-distinctive term, applicable to a great number of individual members of a class or group.

Glossematics: Hjelmslev's term for the study of glossemes.

Glosseme: The smallest meaningful linguistic unit.

Glottal: A sound phonated in the larynx (hence also called laryngal) by a narrowing or closing of the vocal chords.

Grapheme: A minimum unit of writing, not capable of subdivision, e.g. the letter *d* in din, or *g* in dig.

Graphemics: The study of systems of writing and their relationships to linguistic systems (also called graphonomy).

Heteronym: A word identical with another word of the same language as to written form, but different in sound and meaning.

Historical linguistics: Diachronic linguistics.

Homographic spelling: Any system of writing in which the same sign always represents the same sound, and any given sound can be represented by one and the same sign only.

Homonym: A word identical in written form and in sound with another word of the same language, but different in origin and meaning.

Intensional: Conveying intent, connotation, implication or depth of meaning.

Labial: In phonetical terminology, a consonant pronounced with the lips.

Labio-dental: In phonetical terminology, a consonant pronounced with the lower lip touching the upper front teeth.

Labio-velar: In phonetical terminology, a consonant pronounced with the lips rounded and the back part of the tongue arched towards the soft palate.

Langue: The total of engrams of language forms in the brains of the members of a speech community.

Lateral: In phonetical terminology, a consonant pronounced with complete closure in the front of the oral cavity but with incomplete closure at the sides, to permit there the escape of air.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Lexical: Relating to the total stock of linguistic signs or morphemes of a given language.

Lexical form: A linguistic form considered merely in its purely lexical character.

Lexicography: The definition and description of the various meanings of the words of a language or of a special terminology.

Linguistics: The science of language.

Literal translation: A word-for-word translation of an utterance or text into another language.

Literary language: That dialect of a language which is regarded as the best and is used for literary purposes.

Metaphor: A figure of speech in which one word is employed for another.

Metonymy: The substitution of a word for another word having a closely related meaning.

Morpheme: A distinct linguistic form, semantically different from other phonetically similar or identical linguistic forms. The morpheme is the minimal meaningful unit of language; e.g. "boys" consists of two morphemes (boy and *s*), "boyishness" of three (boy, ish, ness).

Morphology: The science and study of the forms and formation of words, i.e. the study of the ways and methods of grouping sounds into sound-complexes, called words, of definite, distinct, conventional meanings.

Nasal: In phonetical terminology, a sound produced with the uvula lowered, allowing the air to escape through the nose, so that the nasal cavity acts as a resonator.

Object: The word or word-group or phrase designating the person or thing at which the action expressed in the sentence is directed.

Opposition: Every phoneme in a language is opposed to every other, since a change from one phoneme to another in a chain of speech should give either a new form or an unintelligible one. Oppositions can be classified as bilateral, multilateral, disjunctive, isolated, proportional or binary.

Palatal: In phonetical terminology, a consonant pronounced with the surface of the tongue arching toward or held near or against the hard palate.

Paradigm: (1) A complete set of all the various forms of a word (conjunction, declension). (2) A model, pattern or example.

Paradigmatic: Relating to or constituting a paradigm. In phonemics, the most important use of this adjective can be summed up as follows: the listing of all the phonemes of a language, in series, orders and isolated elements forms a paradigm, from which the individual phonemes are selected as needed for the chain of speech.

Parent language: The language from which another language or other languages developed is said to be the parent language of the latter.

Parole: French for "word". According to the linguistic terminology introduced by F. de Saussure, *parole* (usually translated into English as "speech") consists in the use of a language by a given individual in a given case, in contradistinction

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

to the *langue* ("language" or "tongue") as its communal and codified aspect.

Person: The grammatical distinction between the speaker, listener and subject of an utterance.

Phantom word: A word that came into being through an error of a lexicographer or printer.

Pharyngal: In phonetical terminology, a consonant phonated at the pharynx.

Philology: In England the same as Linguistics; elsewhere the study of literature.

Phoneme: A minimum unit of distinctive sound-feature. It may be defined also as a minimal bundle of relevant sound features.

Phonemics: The study, analysis and classification of phonemes, their relationships and changes.

Phonetic (noun): In Chinese writing, the symbol within a compound ideograph, which gives a clue to the pronounciation of the spoken word represented by the ideograph.

Phonetic, phonetical (adj.): Pertaining to, derived from, according to, or based on sounds.

Phonetics: The study of the gross physiological and acoustic features of speech-sound.

Plosive: In phonetical terminology, a consonant produced by completely closing the nasal and oral air-passages (implosion) resulting in a retention of air, and then suddenly opening the closure (explosion).

Prosody: The science and art of versification and of the rhythmic structure of sound in speech, especially in poetry. Adj.: prosodic.

Reduplication: The complete or partial repetition of an element or elements.

Reduplicative expression (or phrase): An expression or phrase consisting of two words which are synonyms or near-synonyms for the same idea connected by a conjunction such as "and"; the two words are usually alliterative or rhyming or both alliterative and rhyming.

Related languages: Languages which have developed from the same parent language.

Rhyme: In general, the use of an identical sound or syllable at the ends of adjacent words or words placed at definite intervals.

Rhythm: A harmonical succession of sounds, consisting in a regular periodicity in a series of phonemes, constituting or contributing to the measured movement or musical flow of language.

Root: The ultimate constituent element common to all cognate words, i.e. that element of a word which remains after the removal of all flexional endings, formatives, etc.

Semantics: A science dealing with the relations between linguistic symbols (words, expressions, phrases) and the objects or concepts to which they refer — and with the history and changes in the meanings of words.

Semiology: The English equivalent of the French term "semiologie", introduced by F. de Saussure as the designation of the science and study of signs in general, of

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

which linguistics, as the science and study of linguistic signs, is regarded by him as a mere subdivision.

Spirant: In phonetical terminology, a synonym of "fricative"; a sound produced with some constriction, but with some degree of aperture at the place of articulation.

Spoken chain: The chain or flow of speech in a given situation.

Stress: Special emphasis on a sound or sound-group.

Synchronic linguistics: Linguistic study limited to the recording and analysis of a given linguistic state or stage.

Tagmeme: The smallest meaningful unit of grammatic form.

Taxeme: An individual grammatical feature. The smallest unit of form in grammar.

Unvoiced: A phoneme pronounced without tension of the vocal chords, e.g. *t*.

Variant: (1) A related but not identical form. (The difference may be phonetical, morphological, a matter of meaning or of written appearance.) (2) An alternative form. (3) Often used instead of the term allophone.

Voiced: In phonetical terminology, said of a consonant pronounced with a vibration of the vocal chords.

Voiceless: In phonetical terminology, said of a consonant pronounced without any vibration of the vocal chords.

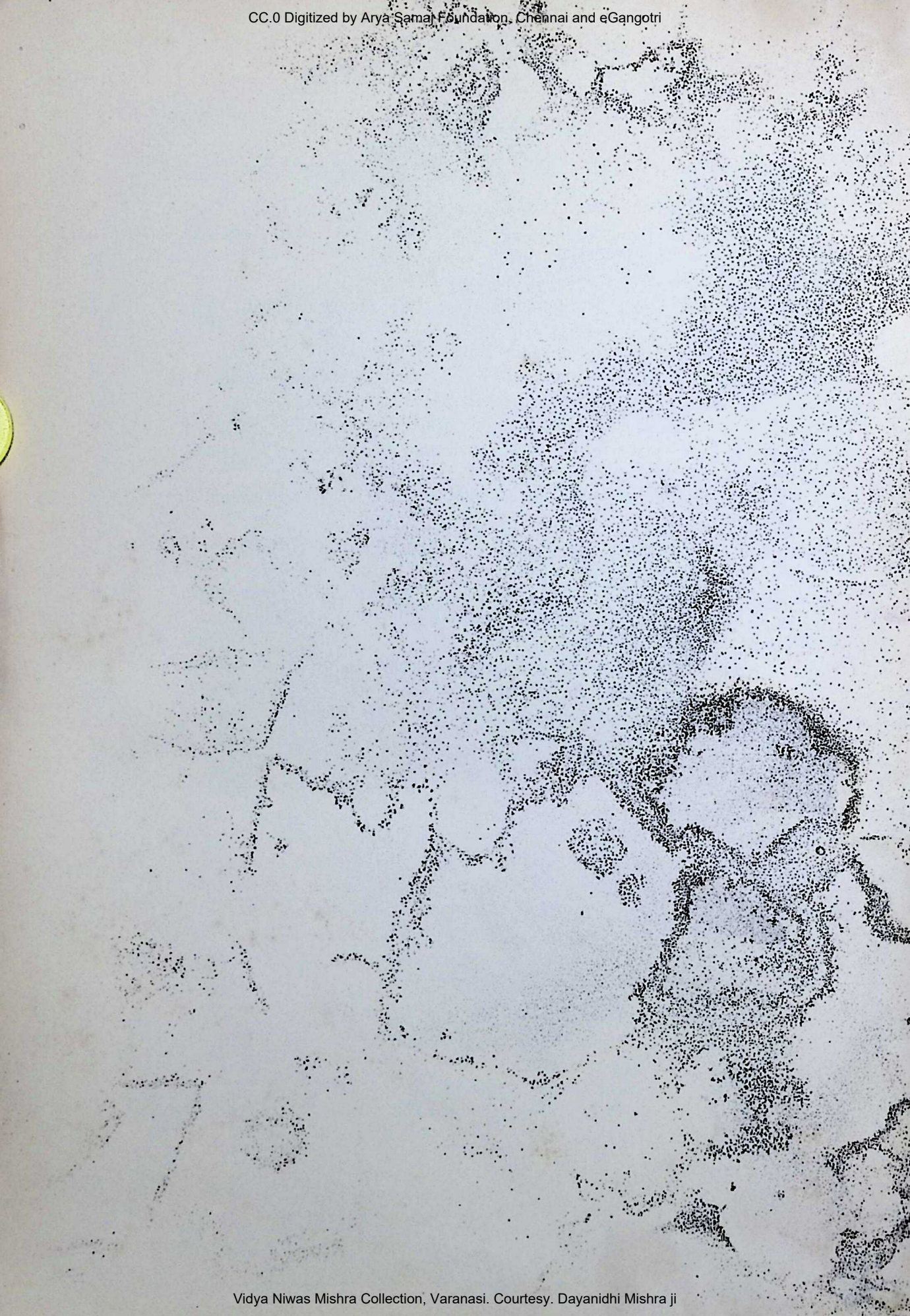
Vowel: (1) A sound produced with a vibration of the vocal chords, by the unobstructed passage of air through the oral cavity. (2) A letter representing such a sound.

Word: A spoken or written symbol of an idea, usually regarded as the smallest independent sense-unit.

Word order: The sequence or relative position of words in a sentence.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri



Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

BIBLIOGRAPHY

The Bibliography of works concerned with mathematical linguistics is quite extensive. In addition to the items listed below the reader is referred to the Bibliography in the author's *Language as Choice and Chance*, to the *Bibliographie Critique de la Statistique Linguistique* by P. Guiraud and J. Whatmough (Utrecht, 1953), and to the forthcoming supplement of the last named work which is to be edited shortly by the Committee on Statistical Linguistics (Comité International Permanent de Linguistes C.I.P.L.); also to the *Bibliography on Information Theory (Communication Theory) — Cybernetics* F. L. H. M. Stumpers, with 3 supplements.

Abbott, E. A., *Flatland, A Romance of Many Dimensions* (1884)

Aitchison, J., and Brown, J. A. C., *The Lognormal Distribution* (Cambridge, 1957).

Anscombe, F. J., "The transformation of Poisson, binomial and negative binomial data", *Biometrika*, 35 (1948), 246–54.

——, "Sampling theory of the negative binomial and logarithmic series distributions", *Biometrika*, 37 (1950), 358–82.

Baker, H. F., *Principles of Geometry*, I (Cambridge, 1922).

Bartlett, M. S., "The square root transformation in the analysis of variance", *J. R. Statist. Soc.*, Suppl. 3 (1936), 68–78.

Bauer, W., *Wörterbuch zum Neuen Testament*, 4. Aufl. (1954).

Bloch, B., and Trager, G. L., *Outline of linguistic analysis* (Baltimore, 1942).

Brain, Sir R., *Aphasia, Apraxia and Agnosia, Neurology*, 2nd edition, by S. A. Kinnier Wilson and A. N. Bruce, Vol. 3 (London, 1955).

Brillouin, L., *Science and Information Theory* (New York, 1956).

Bruder, C. H., *Concordantiae omnium vocum Novi Testamenti*, 7th edition (Göttingen, 1913).

Buch, K. R., "A note on sentence length as random variable", *11. Skandinaviske Matematikerkongress. Proc.* (1949), 272–275.

Capell, A., *The linguistic position of South-Eastern Papua* (Sidney, 1943).

Chadwick, J., "Mycenaean Greek", *Proceedings of the Eighth International Congress of Linguists, Oslo, 1957.*

Chambers, E. G., and Yule, G. U., "Theory and observation in the investigation of accident causation" (including discussion by J. O. Irwin and M. Greenwood), *J. R. Statist. Soc.*, Suppl. 7 (1942), 89–109.

Champernowne, D. G., "A model of income distribution", *Econ. J.*, 63 (1953), 318.

Chotlos, J. W., *Studies in Language Behaviour. Psychol. Monographs*, 56 (1944).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Chrétien, C. D., "Word distribution in southeastern Papua", *Language*, 32 (1956), 88–108.

Corbet, A. S., Fisher, R. A., and Williams, C. B., "The relation between the number of species and the number of individuals in a random sample of an animal population", *J. Anim. Ecol.*, 12 (1943), 42–58.

Court, N. A., "Is mathematics an exact science?", *The Scientific Monthly*, 47 (1948), 119–123.

——, "Plane geometry and plain logic", *The Scientific Monthly*, 83 (1956), 28–34.

——, "The might and plight of reasoning", *Scripta Mathematica*, 22 (1956), 45–56.

Czuber, E., *Wahrscheinlichkeitsrechnung* (Leipzig, 1903), 321.

Darwin, J. H., "Population differences between species growing according to simple birth and death processes", *Biometrika*, 40 (1953), 370.

Davis, Harold T., *The Analysis of Economic Time Series* (Principia Press, 1941).

Devooght, J., "Sur la loi de Zipf-Mandelbrot", *Bull. de la Classe des Sciences, Roy. Academy of Belgium*, No. 4 (1957).

Dewey, G., *Relative Frequency of English Speech Sounds* (Cambridge, Mass., 1923).

Diderichsen, Paul, "The importance of distribution versus other criteria in linguistic analysis", *Proceedings of the Eighth International Congress of Linguists, Oslo, 1957.*

Eddington, A. S., *Space, Time and Gravitation* (Cambridge, 1920).

——, *Fundamental Theory* (Cambridge, 1947).

Einstein, A., *Über die spezielle und allgemeine Relativitätstheorie*, 10th edition (Braunschweig, 1920).

Elderton, W. P., "A few statistics on the length of English words", *J.R.S.S.*, Series A., Vol. 62 (1944), 436–445.

Eldridge, R. C., *Six Thousand Common English Words* (Buffalo, 1911).—Mentioned in Zipf (1949).

Feller, William, *An Introduction to Probability Theory and Its Applications*, I (2nd ed., New York, 1950).

Fletcher, A., Miller, J. C. P., and Rossenhead, L., *An Index of Mathematical Tables* (London, 1946).

Fowler, M., "Herdan's Statistical Parameter and the Frequency of English Phonemes", *Studies presented to Joshua Whatmough on his sixtieth birthday* (The Hague, 1957).

French, N. R., Carter, C. W., Jr., and Koenig, W., Jr., "Words and sounds of telephone conversations", *Bell System Technical Journal*, 9 (1930).

Geden, A., *Concordance of the New Testament*, 3rd edition (1926).

Good, I. J., *Probability and the Weighing of Evidence* (London-New York, 1950), Chapter 6.

——, "Rational decisions", *J. R. Statist. Soc.*, Ser. B., 14 (1952), 107–114.

——, "On the population frequencies of species and the estimation of population parameters", *Biometrika*, 40 (1953), 237–264.

——, *Uncertainty and Business Decisions*, edited by C. F. Carter, G. P. Meredith and G. L. S. Shackle (Liverpool, 1954), Ch. 3, p. 104.

——, "The likelihood ratio test for Markoff chains", *Biometrika*, 42 (1955), 531–3.—Errata in 44 (1957), 301.

——, (jointly with G. H. Toulmin), "The number of new species, and the increase in population coverage, when a sample is increased", *Biometrika*, 43 (1956), 45–63.

——, "Some terminology and notation of information theory", *Proc. Inst. Elec. Eng.*, Part C (3), 103 (1956), 200–204.

——, "The surprise index for the multivariate normal distribution", *Annals Math. Stat.*, 27 (1956), 1130–5.—Errata: Dec. 1957.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

——, "Distribution of word frequencies", *Nature*, 179 (1957), 595.

——, (jointly with R. B. Dawson), "Exact Markov probabilities from oriented linear graphs", *Annals Math. Stat.*, Dec. 1957.

——, (jointly with K. Caj. Doog), "A paradox concerning information theory", *Information and control* (1958).

Goodman, L. A., "On the estimation of the number of classes in a population", *Ann. Math. Statist.*, 20 (1949), 572–9.

Greenberg, J. H., "The measurement of linguistic diversity", *Language*, 32 (1956), 109–115.

Greenwood, M., and Yule, G. U., "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents", *J. R. Statist. Soc.*, 83 (1920), 255–279.

Guiraud, P., *Les Caractères Statistiques du Vocabulaire* (Paris, 1953).

——, "Langage et Communication...", *Bulletin de la Société de Linguistique de Paris*, 50, Pt. 1 (1954), 119–133.

——, "Emprunts et Équilibre phonologique", *Zt. Rom. Philol.*, 74, Pt. 1/2 (1958), 78–88.

Guthrie, D., *The Pastoral Epistles* (London, 1957).

Halle, Morris, "In Defence of the Number Two", *Studies presented to J. Whatmough* (The Hague, 1957).

Hanley, Miles L., *Word Index to James Joyce's Ulysses* (Univ. of Wisconsin Press, 1937).

Harary, F., and Paper, H. H., *Language*, 33 (1957), 143–170.

Hardy, G. H., *Divergent Series* (Oxford, 1949).

Harris, Zellig S., "Co-ordination and transformation in linguistic structure", *Language*, 33 (1957), 283–341.

——, "Discourse Analysis", *Language*, 28, (1952), 1–30.

Harrison, J. L., "Starch products and the insects infesting them, as examples of the logarithmic series", *Ann. Eugen.*, 12 (1945), 280–2.

Harrison, P. N., *The Problem of the Pastoral Epistles* (Oxford, 1921).

Hawkins, J. C., *Horae Synopticae* (Oxford, 1899).

Head, Henry, *Brain*, 34 (1920), 87.

Henschen, S. E., (1920). *Klinische und anatomische Beiträge zur Pathologie des Gehirns*, V, VI and VII (Stockholm, 1920 et seq.).

Herdan, G., *Small Particle Statistics* (Amsterdam, 1953).

——, "Language in the light of the statistical theory of information, II", *Metron, International Review of Statistics*, XVII, 3–4 (Rome, 1955).

——, "A new derivation and interpretation of Yule's 'Characteristic' K", *J. of Applied Mathematics and Physics*, VI (1955), 332.

——, *Language as Choice and Chance* (Groningen, 1956).

——, "Chaucer's authorship of the Equatorie of the Planetis", *Language*, 32 (1956), 254.

——, "The Numerical Expression of Selective Variation in the Vowel-Consonant Sequence in English and Russian", *Studies presented to Joshua Whatmough on his Sixtieth Birthday* (The Hague, 1957).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

——, "An inequality relation between Yule's Characteristic K and Shannon's Entropy H", *J. Appl. Mathem. Phys.* (Zamp), IX, Pt. 1 (1958), 69–73.

——, "The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics", *Biometrika*, 45, Pts. 1/2 (1958), 222–228.

——, "The mathematical relation between Greenberg's index of linguistic diversity and Yule's Characteristic", *Biometrika*, 45, Pts. 1/2 (1958), 268–270.

——, "The mathematical relation between the number of diseases and the number of patients in a community", *J.R.S.S.*, Series A, 120, Pt. 3 (1957), 320–330.

——, "The relation between the functional burdening of phonemes and the frequency of occurrence", *Language and Speech*, I (1958), 8–13.

——, "The Hapax Legomenon: a real or apparent phenomenon?", *Language and Speech*, II (1959), 26–36.

——, "The statistical interpretation of Aphasia", *Confinia Psychiatrica*, I (1958), 143–158.

——, and Grayston, K., "The authorship of the Pastorals in the light of statistical linguistics", *New Testament Studies*, 6 (1959), 1–15.

Hjelmslev, L., *Omkring sprogteoriens grundlaeggelse* (København, 1943). English translation by F. J. Whitefield. *Indiana University Publications* (1953).

Hume, D., *Enquiry concerning human understanding* (London, 1738–40).

Huxley, J. S., Problems of relative growth (London, 1932).

Jackson, Hughlings., "Croonian Lectures", *British Medical Journal*, I (1864), 123.

Jahnke, E., and Emde, F., *Funktionentafeln mit Formeln und Kurven*, 2nd ed. (Leipzig-Berlin, 1933).

Jakobson, R., "Typological studies and their contribution to historical linguistics", *Proceedings of the Eighth Internation Congress of Linguists, Oslo, 1957.*

Jeffreys, H., *Theory of Probability*, 2nd ed. (Oxford, 1948).

——, and Jeffreys, B. S., *Methods of Mathematical Physics* (Cambridge, 1946).

Jespersen, O., *Philosophy of Grammar* (London, 1924).

Johnson, W. E., Appendix to "Probability: deductive and inductive problems", *Mind*, 41 (1932), 421–3.

Jones, E. W., and Williams, C. B., "The index of diversity as applied to ecological problems", *Nature*, 155, (1945), 390.

Joos, M., "Review", *Language*, 12 (1936), 196–201.

Kendall, David G., "On some modes of population growth leading to R. A. Fisher's logarithmic series distribution", *Biometrika*, 35 (1948), 6.

Kendall, M. G., Babington, and Smith, B., "Randomness and random sampling numbers", *J.R.S.S.*, 101 (1938), 147.

Kendall, M. G., and Buckland, W. R., *A Dictionary of Statistical Terms* (Edinburgh, 1957).

Kolmogoroff, A. N., "Теория передачи информации", U.S.S.R. Acad. of Science, Session Oct. 1956 (1957).

Krámský, J., "On the Quantitative Phonemic Analysis of English Mono- and Di-syllables", *Philologica, Cizojazyčná příloha CMF*, VIII (1946).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

——, "Statistiky Pohyb ve Vývoji Angličtiny", *Universitas Carolina, Philologica*, I, No. 1 (1955), 59–65.

Leavens, Dickson, H., *Econometrica*, 21 (1953), 630.

Locke, John, *An Essay Concerning Human Understanding* (London, 1689).

Longacre, R. E., "Items in Context…", *Language*, 34, (1958), 482–491.

Mandelbrot, Benoit, "An informational theory of the statistical structure of language", *Communication Theory*, ed. by Willis Jackson (Butterworth, 1953).

A. A. Markoff, "Примеръ статистическаго изследованія надъ текстомъ "Евгенія Онегина" иллюстр. связь испытаній въ цепь", *Bulletin de l'Académie Imperiale des Sciences de St. Pétersbourg*, 1913.

——, "Изследованіе замечательнаго случая зависимыхъ испытаній."

——, "Объ одномъ случае испытаній связанныхъ въ сложную цепь", *Bulletin* of 1907 and 1911.

Mathesius, V., *Čeština a obecný jazykozpyt* (Prague, 1947).

Mayendorf, N. von, *Die aphasischen Symptome und ihre kortikale Lokalisation* (Leipzig, 1911).

Mendenhall, T. C., "The characteristic curves of composition", *Science*, 9 (supplement 214) (1887), 237–49.

——, "A mechanical solution of a literary problem", *Pop. Sci. Mon.*, 9 (1901), 97–105.

Mersand, Joseph, *Chaucer's Romance Vocabulary* (New York, 1939).

Mises, R. von, *Wahrscheinlichkeitsrechnung* (Leipzig-Wien, 1931).

Morgan, A. de, *A Budget of Paradoxes*, 2nd edition (London, 1915).

Morgenthaler, R., *Statistik des Neutestamentlichen Wortschatzes* (Zürich-Frankfurt am Main, 1958).

Newbold, E. M., "Practical applications of the statistics of repeated events, particularly to industrial accidents" (Including discussion by M. Greenwood, D. R. Wilson, M. Culpin, E. Farmer, and L. Isserlis), *J. R. Statist. Soc.*, 90, (1927), 487–547 (esp. Appendix, pp. 518–535).

Pei, M. A., and Gaynor, F., *A Dictionary of Linguistics* (New York, 1954).

Preston, F. W., "The commonness, and rarity, of species", *Ecology*, 29 (1948), 254–283.

Rapaport, A., Review in *Behavioral Science*, 2, No. 4 (1957).

Reye, Th., *Geometrie der Lage*, 3 vols., 3rd edition (Leipzig, 1886).

Ross, A., "Philological probability problems", *J. R. Statist. Soc.* (B), 12 (1950), 39.

Samuel, Viscount, *Essay in Physics* (Oxford, 1951).

Shannon, C. E., "A mathematical theory of communication", *Bell Syst. Techn. J.*, 27 (1948), 379–423.

Sherman, L. A., *Analytics of Literature* (Boston, 1893).

Simon, H. A., "On a class of skew distribution functions", *Biometrika*, 42 (1955), 425–440.

Simpson, E. H., "Measurement of diversity", *Nature*, 163 (1949), 688.

Spang-Hanssen, H., "Typological and statistical aspects of distribution as a criterion in linguistic analysis", *Proceedings of the Eighth International Congress of Linguists, Oslo, 1957.*

Tatlock, John S. A., and Kennedy, A. G., *A Concordance of the Complete Works of Geoffrey Chaucer* (Washington, 1927).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Thorndike, Edward L., "On the number of words of any given frequency of use", *Psychol. Rec.*, 1 (1937), 399.

——, and Lorge, I., *The Teacher's Word Book of 30,000 Words* (Columbia Univ., 1927).

Titchmarsh, E. C., *The Theory of Functions*, 2nd edition (Oxford, 1939).

Todd, J. A., *Projective and Analytical Geometry* (London, 1947).

Trnka, B., *A Phonological Analysis of Present-Day Standard English. Práce z vědeckých ústavů*, XXXVII (Prague, 1935).

Trubetzkoy, N. S., *Anleitung zu phonol. Beschreibungen*, C.L.P. (Brno, 1935).

——, Essai d'une Théorie des oppositions Phonologiques. *J. Psych.*, 33 (Paris, 1936), 5–18.

——, *Grundzüge der Phonologie* (Prague, 1939).

Uspensky, J. V., *Introduction to Mathematical Probability* (New York, 1937).

Vander-Beke, G. E., *French Word Book* (New York, 1929).

Wake, W. C., "Sentence length distributions of Greek authors", *J.R.S.S.*, Ser. A., 120 (1957), 331–346.

Weinreich, Uriel, "Research Frontiers in Bilingualism Studies", *Proceedings of the Eight International Congress of Linguists, Oslo, 1957.*

Wells, R., "Is a Structural Treatment of Meaning Possible?", *Proceedings of the Eighth International Congress of Linguists, Oslo, 1957.*

Whatmough, J., *Language* (London, 1955).

——, *Poetic, Scientific and Other Discourse. Sather Lectures.* (Berkeley–Los Angeles, 1956).

——, "Mathematical Linguistics", *Proceedings of the Eight International Congress of Linguists, Oslo, 1957.*

Whittaker, E. T., and Robinson, G., *The Calculus of Observations*, 4th ed. (London-Glasgow, 1944).
——, and Watson, G. N., *A Course of Modern Analysis*, 4th ed. (Cambridge, 1935).

Williams, C. B., "A note on the statistical analysis of sentence length as a criterion of literary style", *Biometrika*, 31 (1940), 356–361.

——, "Area and number of species", *Nature*, 152 (1943), 264.

——, "A note on the relative size of genera in the classification of animals and plants", *Proc. Linn. Soc.*, Series 162, II (1949–50), 171–175.

——, "The number of publications written by biologists", *Ann. Eugen.*, 12 (1944), 143–6.

——, "Some applications of the logarithmic series and the index of diversity to ecological problems", *J. ecol.*, 32 (1944), 1–44.

——, "Recent light trap catches of Lepidoptera in U.S.A., analysed in relation to the logarithmic series and the index of diversity", *Ann. Eug. Soc. Amer.*, 38 (1945), 357–364.

——, "Yule's 'Characteristic' and the 'Index of Diversity'", *Nature*, 157 (1946), 482.

——, "The logarithmic series and the comparison of island floras", *J. Linn. Soc. Lond.* (in the press).

——, *The generic relations of species in small ecological communities* (1947).

——, "Statistics as an aid to literary studies", *Penguin Science News*, No. 24 (1952), 99–106.

Yule, G. Udny, "A mathematical theory of variation, based on the conclusions of Dr. J. C. Willis, F. R. S.", *Philosophical Transactions, Royal Society of London*, Series B., 213 (1924), 21–87.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

——, "On sentence-length as a statistical characteristic of style in prose", *Biometrika*, 30 (1939), 363–84.

——, *The Statistical Study of Literary Vocabulary* (Cambridge, 1944).

——, *A Statistical Study of Vocabulary* (Cambridge, 1944).

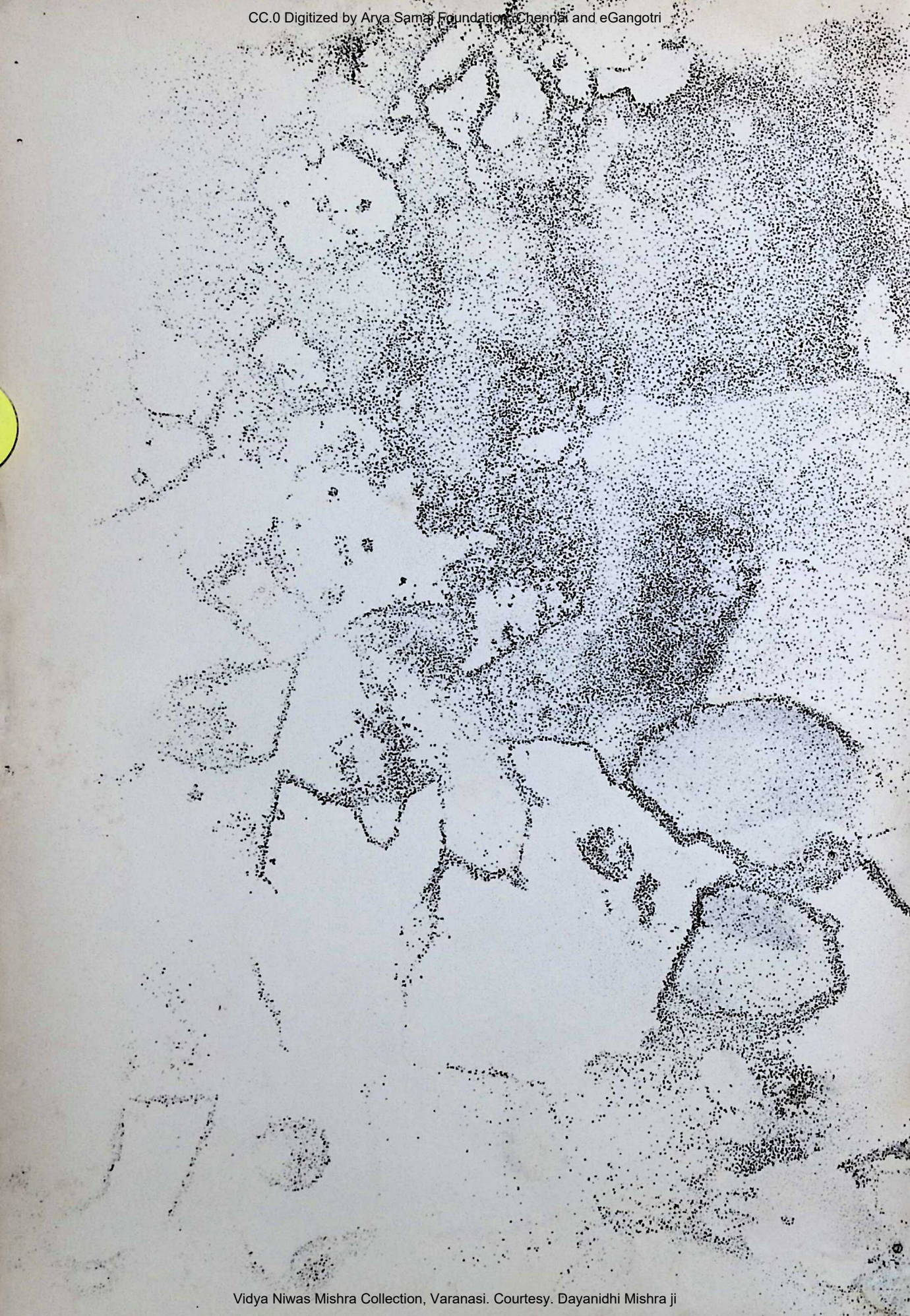
—— and Kendall, M. G., *An Introduction to the Theory of Statistics*, 11th ed. (London, 1941).

Zipf, G. K., *Selected Studies of the Principle of Relative Frequency in Language* (Cambridge, Mass., 1932).

——, *Human Behaviour and the Principle of Least Effort* (Cambridge, Mass., 1949).

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri



Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

INDEX

Items in the Glossaries and the Bibliography, which are already in alphabetical order, are not listed.

Abbott, E. A., 240
Additiveness of information, 178, 179
Aelfric Homilies, word count of, 89
Aitchison, J., 133, 333ff.
Alphabetical distribution of nouns according to etymology, 161, 162
Analysis of variance, 318, 320; principle of evaluation, 322
Antisymmetry, 125
Aphasia, 239
Area-volume relation, 285
Aristotle, 324
Arithmetic probability paper, 43
Artificial codes, 182
Authorized version, 267
Average gap length, 259

Babington Smith, B., 115
Beethoven, 229
Belevitch, V., 37
Bernoullian correlation, 96
Bilogarithmic growth-rate as style characteristic, 29, 31
Bilogarithmic straight line, 207; law, 217
Bilogarithmic Type/Token ratio, extension to different texts, 28
Binary coding, 314
Binary number system, 174
Books of the New Testament, their vocabulary connection, 104
Boole, G., 17, 173, 323
Boolean algebra, 173, 324
Boole's law of duality, 323, 324
Born, M., 305
Bose-Einstein statistics, 304
Brain functions, bilateral and unilateral, 234
Broglie, de, 259,
Brown, J. A. C., 133, 333ff.
Brown, R. W., 323
Bühler, K., 118
Bunyan, works of, 60
Butler, S., 261

Capell, A., 98
Catholic letters, 246
Cause and effect in language, 82
Chance in language, 39
Characteristic *K*, 83; and Entropy H, inequality relation between, 192, 303; as estimate of repeat rate, 94, 337, 338
Chechen, 120, 121
Chinese dictionary structure, 219, 224
Chi-square, computation of, 108
Choice in language, 40
Chomsky, N., 40
Ghotlos, J., 26
Classical statistics, 281
Classification, 211
Classifications of oppositions, 126
Codes, plurality of, 273
Coding, binary, principle of, 301; criterion of quality, 186; symbols fewer than objects, 154
Common error term, 261
Completeness, 123
Compound Poisson law, 77ff.
Concatenation of contexts, 59
Concentration, measure of, 48ff.
Concept content vs. concept size, 220; division, [212
Conceptual structure, 209
Content, statistics suitable for, 281
Contradiction, principle of, 324
Conventional measures of vocabulary correlation, shortcomings of, 97
Correlation between oppositions, 126
Correlation coefficient and Chi-square, relation between, 100
Cost of transmission of information, 37, 171
C-plane, of content, 256
Cross-classification of oppositions by Trubetzkoy, 126

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

Cross-word puzzle, 240

Devooght, J., 26
Dewey, G., 208
Dialectic movement of history, 331
Dictionary and occurrence distributions, relation between, 130
Dimension, 327
Dirr, A., 118
Discourse analysis, statistical version of, 59
Distinctive function, 316
Distribution of loanwords in French according to origin, 166, 167; of vocabulary, 319
Distributionalists vs. Prague school, 240
Dostoevsky, F., 273
Double relativity of phoneme counts, 123, 127, 132
Doubling mechanism of language, 235, 238
Drama vs. fiction, 270
Duality, 243; in higher dimensions, 243, 271; law of, 173; linguistic principle of, 328; as tool of research, 239
'Duplication', 263, 264
Dyadic, (dual) number system, 174; structure, four-fold root of, 311

Eddington, Sir A. S., 269, 331
'Egg' of Columbus, 244
'Eigencontrol' in poetic diction, 278
'Eigenstatistics' of language, 314
Einstein, A., 15, 82
Ektropy, 171
Elderton, 183
Elton, O., 142
English consonant phonemes, dictionary distribution of, 128; running text count of, 129
Entropy, definition of, 176, 302; linguistic interpretation of, 171; and linguistics, 192; as a measure of information, 177; transformation of, 192
Ephesians, vocabulary connectivity of, 110
Epistomology, 314
E-plane, of expression, 256
Equivalent-chain relation, 59
Ergodicity of word masses, 295
Estimators of relative areas, 282, 283
Expectation of probability, 326
Experimental design, 318
Expression, statistics suitable for, 281

Fibonacci numbers, 191
Fiction, craft of, 270
Flatland, 240ff.
Fowler, M., 90
Frequency of linguistic forms, 21ff.
Functional burdening, 34; of phonemes, 121, 127; and utilisation of phonemes in German, 122; of words with meaning, 119
Functional relations in mathematics, 34
Gap frequency, 300
Gaps, between repetitions, distribution of, 154; between words, chance distribution of, 155
Genera and species numbers, 207
Geometrical duality, 230, 231; in four dimensions, 257
Geometrical probability, 282
'Geometrie der Lage', 311
Geometries in linguistics, 244
Geometry, in space, 310; in structural linguistics, 14
Gini, G., 333, 334
'Gleichschaltung', 263, 264
Glinz, H., 41
Goethe, 277; Essay on Nature, 263, 268
Good, I. J., 36, 83, 87, 185
Grammar style characteristics, 293
Grayston, K., 30
Greek New Testament, vocabulary and occurrences, 23; word count of particular books, 24, 25
Greenberg, J. H., 92
Greenberg's index of linguistic diversity, 93
Group wave, 259
Guiraud, P., 22, 33, 68, 164
Guthrie, D., 30

Hahn, E. A., 239
Halle, M., 40
Hapax legomenon, definition of, 66; real or apparent, 66
Harary, F., 123
Harnack, A. v., 25
Harris, Z., 59, 238
Harrison, P. N., 25, 30
Hawkins, J. C., 25
Haydn, 229
Hegel, 331
Herdan, G., 26, 30, 83, 87, 129, 184, 185 n., 208, 239, 307
Hjelmslev, L., 210
Hoenigswald, H. M., 70

Identity of text samples, 250
Incomplete Balanced Block Design, 317
Incomplete periodic pattern of words, 260
Independence, of sign and meaning, 295, 296; of word form from word content, 39
Inequality, between K and H, 193; applications of, 195, 303
Information, linguistic interpretation of, 171; rate of transmission of, 175; storage of, 179; theory, 14; units of, 173
Initials, repeat rate of, 162
Intellectual activity, 317

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

'Iphigenie', 277

Jakobson, R., 314, 325
James, Henry, 273
Japanese, conservative and innovating, 125
Jensen, J. B., 56
John, writings of, 32
Johnson, Dr., 39
Joos, M., 36

K as coefficient of variation of the mean, 88
Kant, 314
Kendall, M. G., 115, 295; et al., 295
Kolmogoroff-Smirnoff Test, 56
Koutsoudas, A., 37, 152
Kramsky, J., 88, 127, 132
Kraus, K., 230, 261, 264, 278
'Kreutzer Sonata', 275

'Language, as choice and chance', 21, 34, 42, 43, 70, 83, 86, 88, 96, 133, 186, 195, 196, 239, 287, 294, 295, 323; as creating thought, 230; decay of, 263
Language-conditioned vs. style-conditioned characteristics, 117
Language Contact, statistical analysis of, 158
Language mixture, 158
Language properties, non-measurable, 308
Language statistics, different types of, 115; objections against, 117; sources of, 21
Language structure, 210; mathematical characteristics of, 281
Language translation, 196; conditioned repeat rates in, 300
Latin legal maxim, 261
Laws of chance in linguistics, advantage of establishing, 42
Law of duality, statistical form of, 326
Least squares, sum of, 298, 299
Letter occurrence, connected and disconnected, [143
Lexian universe, 293
Lexico-statistics, 123
Lexis' Q, 145, 151, 254; error of, 145
Linguistic control, 316
Linguistic duality, 15; abuse of, 263; epistimological nature of, 314; four-fold root of, 330
Linguistic relationship, established by Chi-square, 99; special vs. general components, 100
Linguistic sampling, 282
Linguistic theorems, duality of, 232, 233
Linguists, International Congress, Oslo 1957, 26
Linkage between vowels and consonants, 145
Literary and linguistic structure, 271
Loan-words in French, 164
Logarithmic probability paper, 43
Logarithmic series, 307
Logarithmic transformation, 307
Logarithmic Type/Token ratio, 26
Lognormal distribution, 42ff., 46ff., 333; moment-property of, 333
Lognormal law, 43
Lognormal variable, cumulative distribution of, 43
Lognormality of linguistic distributions, 133
Longacre, R. E., 59
Lorenz diagram, 48, 49
Lorenz factor, 334; as function of standard deviation, 335; and logarithmic standard deviation, 50
Lubbock, P., 270, 272
'Lumpers', 211
Luther, M., 267

Macaulay, essays of 60; Essay on Bacon, 51, 52; History of England, 55
Mandelbrot, B., 33, 36, 37
Mark 16, 9-20, 287, 291
Markoff, A. A., 141
Markoff chain, 141, 196; complex, 147; mechanism, 146; simple, 147; between translational equivalents, 198
Marx, K., 331
Mathematical linguistics, main division of subject, 16
Mathematical philology, 15, 262
Mathematics, types of — used in mathematical linguistics, 14
Mathematics of measurement, 262
Mathematics of order and position, 262
Mathesius, G., 121
Maxwell-Boltzmann statistics, 305
Mean difference, coefficient of, 333
Meaning, segmentation of, 210
Mechanical translation and Markoff chain, 152
Melanesian languages, 98
Metaphysics, 314
Mises, R. v., 115
Mises' principle, objections against, 115
Moment distributions of lognormal variable, 46, 134; of word length, 136
Moments and moment distributions, 43
Morgenthaler, R., 22ff., 204, 246
Motor (sensory) activity, bi-laterality of, 317
Mozart, 229
Music, ways of interpreting, 229
Mysticism, 314

Nazi German, linguistic analysis of, 263
Nazi mentality, 264
Negative binomial, 207
'New Statistics', 282
New Testament, books, 249; Type/Token ratio

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

of individual books, 31; vocabulary distribution, 245; word count, 44
'Noise', 196
Normal or Gaussian law, 42
Normal law, cumulative distribution of, 43
Number statistics, 281
Numerical table for calculating numbers of genera and species, 340, 341; for calculation of Entropy *H*, 338-340; of random partitioning function, 341 ff.

Occurrence and Lexicon distribution, information values of, 138
Oettinger, K., 314
Omar Khayam, 184
Opposition, contradictory vs. contrary, 325, 326
Oppositions, one and more dimensional, 126; proportional and isolated, 126
Optimal coding, condition of, 184; and Fibonacci numbers, 189; generalisation of condition for, 185; mathematical structure of, 186, 336
Optimal repeat rate, 195

Paper, H. H., 123
Parameters and methods, specificity of, 311
Pastoral Letters, 29, 246
Pastorals, vocabulary connectivity of, 111
Pauline Epistles, their vocabulary connection, 106; Type/Token ratio of individual letters, 30
Pauline letters, 248
Periodicity of words, 258
Philological duality, 272
Phoneme count in dictionary vs. phoneme count in running texts, 123
Phonemes, functional burdening and utilisation, 121; observed vs. possible frequency of, 119; vs. phonetic variants, 116
Phonemic channel vs. letter channel, 139
Phonetics vs. phonology, 260
Phonological equilibrium, 165
Phonological opposition, 315
Phonological oppositions, frequency distribution of, 126
Physical statistics, 281
Pitman's Shorthand, 182
Pitman's Shorthand outline, length of, 183
Plane geometry, 309
Poetic diction and vocabulary connectivity, 260
Poetic diction, linguistic duality of, 278
Poisson law, 77, 207, 254
Population probabilities in language statistics, 292, 294
Primal vs. punctual analysis, 258
Prime Number Theorem, 205
Printed English, information of, 180
'Prinzip des ausgeschlossenen Spielsystems', 115
Probability mechanism of concept creation, 213
Product-moment correlation, 96
Prosodic properties of prose, 255, 259, 260
Punctual vs. primal analysis, 258
Pushkin, A., 'The Captain's Daughter', 374; Evgeny Onegin, 142
Pythagoras, 324

Quantitative language structure, 225

Random partitioning, experimental demonstration of, 64; of vocabulary, 252, 253, 254
Random sampling as operational concept, 115
Rank distribution of frequency according to, 203
Rapoport, A., 325
Reciprocity relation in language, 211
Reciprocity relations, artificial, 202; of linguistic forms, 201; natural, 206
Rectilinearity on a bilogarithmic grid, 31, 78
Redundancy, linguistic interpretation of, 172; in secret codes, 172
Relative (differential) growth in biology, 26
Relative growth of literary products, 27
Relative richness of vocabulary, 75
Relativity of vocabulary ratio, 81; application to dating of literary texts, 73
Repeat rate, 297, 314, 326; bivariate form, 337; geometrical interpretation of, 83; of language forms, 137, 192; as limiting factor, 330; restrictive condition of geometrical interpretation, 86
Repeat rates, conditioned, 300
Repeated occurrences in linear sequences, 197
Repetitiveness of linguistic forms, 154
Reproducibility of proportions, 286
Romance vocabulary in Chaucer, 71; theoretical derivation, 75
Romance vocabulary ratio, its dependence upon text length, 70, 71
Romans and Corinthians, 108
Romans and Galatians, 108
Romans and Pastorals, 109
Rosin-Rammler law, 214, 222
Ross, A. C. S., 39, 81

Sampling counts, 282
Sampling efficiency, 291
Sampling of linguistic data, 292, 293
Sampling in linguistics, 115
Samuel, H., 82
Satire, craft of, 263, 265, 266
de Saussure, F., 17, 34, 39, 327
'Der Schimmelreiter', 276
Schopenhauer, A., 211, 229, 278
Schrödinger, E., 196

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri

Self-duality, 231
Selective variation, numerical expression of, 141
Self-sufficiency of language, 318
Sentence length, and authorship, 55; in Greek authors, 56; language-conditioned elements of, 56; as statistical variable, 54
Septuagint, 25
Sermon on the Mount, 266, 267
Shannon, C. E., 172
Shaw, G. B., 152, 229
Sherman, A., 55
Significance testing in stylostatistics, 287
Significant differences, 317
'Signifiant', geometrical form of, 327
Signifiant and Signifié, independence between, 39
Signs, frequency (use) of, 329
Signs in language, number of, 329
Simon, H. A., 201, 207
Simon's assumptions, inadequacy of, 208
Simon's stochastic model, 116, 208
Skew distributions of statistical linguistics, 201
Shakes, genera and species numbers, 219, 224
Spang-Hanssen, H., 155
Speech function, unilaterality, 233
'Splitters', 211; vs. 'lumpers', 211
Standard deviation according to Bernoullian law, 145
Statistical frequency dictionary, 22
Statistical interpretation of de Saussure's dichotomy of language, 34
Statistical invariant, structure of meaning as, 320
Statistics of language structure, peculiarity of, 13, 14, 281
Statistical linguistics, 13
Statistical regularities, 34
St. John's Gospel, 267
Stochastic dependence between translational equivalents, 199
Storm, Th., 276
Structure of meaning, 249, 251, 252; statistics of, 303
Sub-normal dispersion, 145
Super-normal dispersion, 145
Symbol length, distribution of frequency according to, 202
Symmetry, 125

Taxonomic time, 212
Taxonomic universe, 293
Text, its relation to Type and Token, 15
Theory of sets, 123, 309
Thorndike, E. L., 22
Token, 15, 231
Tolstoy, L., 199, 275
Translational equivalents, 197
Trnka, B., 121, 127
Trubetzkoy, N. S., 22, 120, 127, 141, 314
Truncating the frequency distribution, its effect upon K, 90
Type, 15, 231
Type-content relation, 257
Types of statistics, their blending in language statistics, 14
Type-text duality, 243
Type/Token duality, 229
Type and token, interchangeability, 230
Type/Token ratio, 26
Type-token-text, 244
Typology of languages, 151

Uncertainty relation, 186
Uniformity vs. diversity, 322
Uniqueness in the statistical sense, 86
Unity of content vs. unity of style, 245
Universe of content, 292
Universes of literary statistics, 292
Use of words, choice and chance in, 39; statistical hypothesis of, 38

V, vocabulary partition constant, 250, 251, 307
Van der Beke, 22
Variance of K, 84
ν_m as characteristic of phoneme distribution, 90
ν_m and repeat rate, 86
ν_m, stability in translations, 88
Vocabulary, partitioning of, 245; random partitioning of, 62; richness of, 68
Vocabulary code vs. occurrences, 139
Vocabulary connectivity, 104; vs. random partitioning, 109
Vocabulary correlation, conventional measures of, 96
Vocabulary and occurrence statistics, 117
Vocabulary partitioning, dynamic view of, 255
Vocabulary structure, features of, 61; Incomplete Balanced Block Design of, 318, 320
Vocabulary universe, 293
Vowel-consonant sequence in Russian and English, 150
Vries, M. de, 84

Wake, W. C., 56
'War and Peace', 199, 275
Weaver, W., 172
Wells, R., 252, 256
'West-Östlicher Divan', 277
Whatmough, J., 22, 66 n.
Williams, C. B., 83, 133
Willis, C., 207
Word count distribution, estimation of, 50; lognormality of, 42ff.
Word length, 37; distribution of in dictionary

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

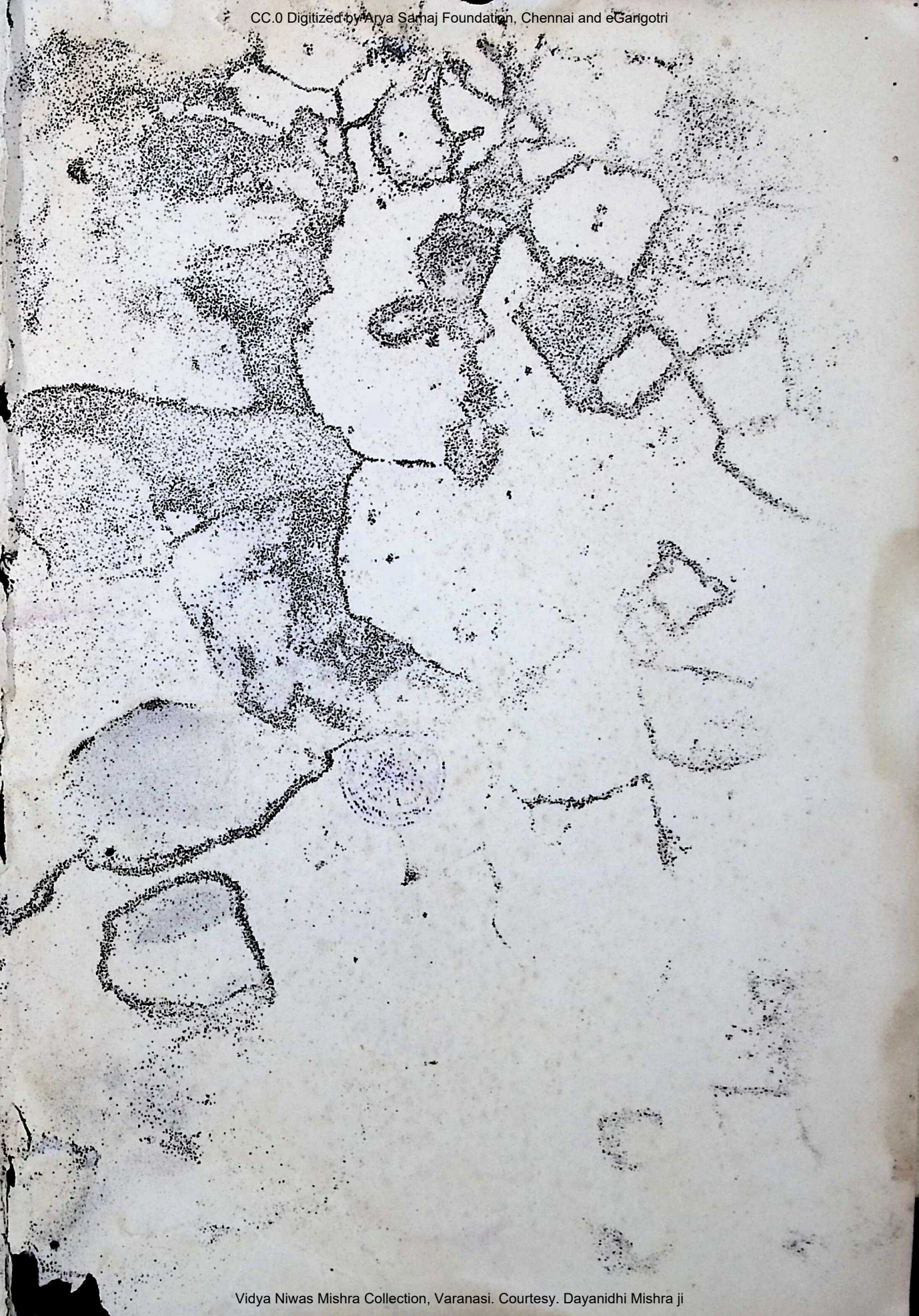
CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

and texts, 133; in German, 118; as lognormal variate, 133
Word masses, mechanical handling of, 204
Words, binary coding of, 188
Yule, G. U., 50, 75, 208
Yule's Characteristic, 93; as estimate of repeat rate, 84; as referring to form not content, 40
Yule, treasure chest simile, 38; word count, 207
Zipf distribution, 207
Zipf law, 33; a mathematical triviality, 33; critical examination of, 35; probabilistic model of, 37

Vidya Niwas Mishra
Sanskrit Department
University of Gorakhpur,
GORAKHPUR.

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation, Chennai and eGangotri



Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji

CC.0 Digitized by Arya Samaj Foundation Chennai and eGangotri

Vidya Niwas Mishra Collection, Varanasi. Courtesy. Dayanidhi Mishra ji